Contents

1 Introduction 2

2 Deriving inference error from the statistical physics of disordered systems 2
  2.1 Replica equations at finite temperature .................................................... 2
  2.2 Replica equations in the low temperature limit ........................................... 5
  2.3 Moreau envelope formulation of the replica equations ................................. 7
  2.4 Accuracy of inference from the perspective of noise ................................... 8
  2.5 Noise-free replica equations ....................................................................... 8

3 Analytic example: quadratic loss and regularization 10
  3.1 General quadratic MSE .............................................................................. 10
  3.2 Optimal quadratic MSE ............................................................................. 11
  3.3 Training and test errors .............................................................................. 12
  3.4 Limiting cases and scaling ......................................................................... 12
  3.5 Quadratic inference through a random matrix theory perspective ................. 13

4 Lower bounds on the error of any convex inference procedure 14
  4.1 Unregularized inference ............................................................................. 14
  4.2 Regularized inference ................................................................................. 15
  4.3 Lower bound on noise-free inference .......................................................... 17
  4.4 Example: bounds on compressed sensing ................................................. 18

5 Optimal inference 19
  5.1 Unregularized inference ............................................................................. 19
  5.2 Regularized inference ................................................................................. 22
  5.3 Analytic bounds on MSE ........................................................................... 25
  5.4 Limiting cases ............................................................................................. 25
  5.5 Optimal noise-free inference ...................................................................... 27
  5.6 The limits of small and large smoothing of Moreau envelopes ................. 27
  5.7 Comparison of optimal M-estimation with MMSE accuracy ...................... 29

A Review of Single Parameter Inference 30
  A.1 Single parameter M-estimation .................................................................. 30
  A.2 Bounds on single parameter M-estimation ............................................... 30
  A.3 Asymptotic analysis of single parameter Bayesian MMSE estimation ........ 31

B Fisher Information 32
  B.1 Fisher information for scalar inference ...................................................... 32
  B.2 Convolutional Fisher inequality .................................................................. 32
  B.3 Fisher information is minimized by a gaussian distribution ....................... 32
  B.4 Relation between MMSE and Fisher information ...................................... 33
  B.5 Cramer-Rao bound .................................................................................... 33

C Moreau envelope and proximal map 34
  C.1 Relation between proximal map and Moreau envelope ............................. 34
  C.2 Inverse of the Moreau envelope ............................................................... 34
  C.3 Conditions for convexity of optimal M-estimators .................................... 35
1 Introduction

Here we provide both detailed derivations of results in the main paper as well as background information about classical results in statistics, to place our results on modern high dimensional inference in perspective.

2 Deriving inference error from the statistical physics of disordered systems

In this section, we derive consistency equations governing the mean squared error $q_s$ between the true parameters $s^0 \in \mathbb{R}^P$ and our estimate of the parameters $\hat{s} \in \mathbb{R}^P$. Einstein summation convention is assumed unless otherwise stated, and measurements $(y_1, ..., y_N)$ are drawn as

$$y_\mu = X_{\mu j} s^0_j + \epsilon_\mu. \quad (1)$$

Given knowledge of the measurements $\mathbf{y} \in \mathbb{R}^N$ and the measurement or input matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ (the $\mu$'th row of $\mathbf{X}$ is the measurement vector $x^\mu$ in the main paper), but not the noise realizations $\epsilon_\mu$, which are drawn iid according to a probability density function $P_\epsilon$, we consider M-estimators of the form:

$$\hat{s} = \arg \min_s \left[ \sum_{\mu=1}^N \rho(y_\mu - X_{\mu j}s_j) + \sum_{j=1}^P \sigma(s_j) \right]. \quad (2)$$

The loss and regularizer $\rho$ and $\sigma$ will be assumed to be convex functions throughout this paper. In what follows we let $\mathbf{X} \in \mathbb{R}^{N \times P}$ be a random matrix with elements drawn iid from a normal distribution with variance $\frac{1}{P}$, as $X_{\mu j} \sim \mathcal{N}(0, \frac{1}{P})$, where the $\frac{1}{P}$ scaling serves to maintain an approximately unit norm for each measurement vector $x^\mu$ so that both the noise $\epsilon$ and signal $\mathbf{X} s^0$ vectors have norms that are $O(N)$ if each component of the noise and signal are $O(1)$.

2.1 Replica equations at finite temperature

**Result 1. Finite temperature RS relations** We let $\beta$ be the inverse temperature and derive the RS (replica symmetric) equations in the high dimensional asymptotic (big-data) limit where $N, P \to \infty$ and the ratio of samples to variables has a finite limit $\frac{N}{P} \to \alpha$. We find that average squared residual error of the energy minimization problem (4), reduces to a set of 4 equations relating scalar order parameters $q_d, q_s, \Lambda_\rho, \Lambda_\sigma$.

$$\frac{q_d}{\Lambda_\sigma} = \frac{\alpha}{\Lambda_\rho} \left( \langle \langle (d - q_d)^2 \rangle_{\rho} \rangle_{\epsilon, z_\epsilon} \right), \quad q_s = \left( \langle \langle s - s^0 \rangle_{\rho} \rangle_{s^0, z_s} \right), \quad \Lambda_\rho = \left( \langle \langle (d)^2 \rangle_{\rho} \rangle_{s^0, z_s} \right)$$

$$P_\rho(d|q_d) \propto e^{-\frac{(d - q_d)^2}{2\rho}}, \quad P_\sigma(s|q_d) \propto e^{-\frac{1}{\alpha P_\rho}(s - q_d s^0)^2 - \beta \sigma(s)} \quad (3)$$

$z_\epsilon, z_s$ are independent random unit gaussians, and $s^0, \epsilon$ are random variables drawn from the parameter and noise probability densities $P_\rho$ and $P_\sigma$ respectively, and we define the effective noise $\epsilon_{s^0} = \epsilon + \sqrt{q_d} z_\epsilon$ and effective parameters $s^0 = s^0 + \sqrt{q_d} z_s$. We define thermal fluctuations from the mean $\delta s = s - \langle s \rangle$ and $\delta d = d - \langle d \rangle$.

**Derivation of Result 1.** We now introduce a physical system which reduces to the above inference problem in the zero temperature limit. The energy function of such a system is

$$E(u) = \sum_{\mu=1}^N \rho(\epsilon_\mu - X_{\mu j} u_j) + \sum_{j=1}^P \sigma(s^0_j - u_j), \quad (4)$$

where we have made a change of variable to write the problem in terms of the residual error $u$ defined as $u = s^0 - s$. The Gibbs distribution on the residual error $u$ is $P_G(u) = e^{-\beta E(u)}$. We then compute the average Free Energy of the system by applying the replica trick based on the identity $\lim_{n \to 0} \frac{1}{n} \left( \langle \langle Z^n \rangle \rangle_{\epsilon} = \langle \langle \log(Z) \rangle \rangle \right)$. Note that $\mathbf{X}, s^0, \epsilon$ here play the role of quenched disorder. To proceed with the replica method, we first compute:

$$\langle \langle Z^n \rangle \rangle_{\epsilon, X, s^0} = \left( \left( \prod_{a=1}^n \right) \int \prod_{\mu=1}^N \text{d}u^a e^{-\frac{1}{2} \sum_{\mu=1}^N \left( \sum_{j=1}^P \rho(X_{\mu j} u^a_j + \epsilon_j) + \sum_{j=1}^P \sigma(s^0_j - u^a_j) \right)} \right)_{\epsilon, X, s^0} \quad (5)$$

Note that, since $X_{\mu j} \sim \mathcal{N}(0, \frac{1}{P})$, for fixed $u^a$, $b^a = \sum_{j=1}^P X_{\mu j} u^a_j$ is a gaussian with a covariance structure $\langle \langle b^a_{\mu j} b^b_{\nu \nu'} \rangle \rangle_{X} = Q_{ab} b_{\mu \nu}$ so that
\[
\langle \langle Z^n \rangle \rangle_{e, X, s^0} = \int \mathcal{D} s^0 P_a(s^0) \int \prod_{ab} dQ_{ab} \int \prod_{a} du^a \prod_{ab} \delta(u^a \cdot u^b - PQ_{ab}) e^{-\beta \sum_{j=1}^{p} \sigma(s_j^0 - u_j^a)} e^{-\beta \sum_{a} u^a - PQ_{ab}},
\]
(6)

where \( |Q| \) denotes the determinant of \( Q \). By noting that part of the integral factorizes in \( \mu \) we have
\[
\langle \langle Z^n \rangle \rangle_{e, X, s^0} = \int \prod_{ab} dQ_{ab} d\hat{Q}_{ab} e^{-\beta \sum_{ab} P \hat{Q}_{ab} Q_{ab}} \int d\eta(E) \prod_{ab} du^a \prod_{a} \delta(u^a \cdot u^b - PQ_{ab}),
\]
(7)

where \( E(Q) = -\log \left( \int d\epsilon P_a(\epsilon) \prod_{a} du^a e^{-\frac{1}{2} \beta^2 \sum_{a} Q_{ab}^{-1} u^a u^b + \frac{1}{2} \beta \sum_{ab} \hat{Q}_{ab} u^a u^b} \right). \)
(8)

We then replace our delta functions with integrals over dummy variables \( \hat{Q} \) on the imaginary axis
\[
\langle \langle Z^n \rangle \rangle_{e, X, s^0} = \int \prod_{ab} dQ_{ab} d\hat{Q}_{ab} e^{-\beta \sum_{ab} P \hat{Q}_{ab} Q_{ab}} \int d\eta(E) \prod_{ab} du^a \prod_{a} \delta(u^a \cdot u^b - PQ_{ab}),
\]
(9)

We can simplify this by noting the factorization of the RHS to find
\[
\langle \langle Z^n \rangle \rangle_{e, X, s^0} = \int \prod_{ab} dQ_{ab} d\hat{Q}_{ab} e^{-\beta \sum_{ab} P \hat{Q}_{ab} Q_{ab}} \int d\eta(E) \prod_{ab} du^a \prod_{a} \delta(u^a \cdot u^b - PQ_{ab}),
\]
(10)

where
\[
S(\hat{Q}) = \log \left( \int d\eta(E) \prod_{ab} du^a \prod_{a} \delta(u^a \cdot u^b - PQ_{ab}) \right).
\]
(11)

We now apply the replica symmetric (RS) assumption about the form of \( Q \), and a corresponding form on \( \hat{Q} \). For \( a \neq b \), \( Q_{ab} = Q_{0}, Q_{aa} = Q_{1}, Q_{ab} = \hat{Q}_{0}, \) and \( Q_{aa} = \hat{Q}_{1} \). It is also useful to define the differences between the replicas \( \Delta = Q_1 - Q_0 \) and \( \hat{\Delta} = \hat{Q}_1 - \hat{Q}_0 \). The RS assumption implies that
\[
b^a = \sqrt{\Delta} x^a + \sqrt{\hat{Q}_0} z,
\]
(12)

where \( x^a, z \) are unit gaussian random variables and \( \Delta = Q_1 - Q_0, \) so that
\[
e^{-E(\hat{Q})} = \left\langle \left( \int d\eta(E) \prod_{a=1}^{n} D x^{a} e^{-\beta \sum_{a=1}^{n} \rho(\epsilon + \sqrt{\Delta} x^a + \sqrt{\hat{Q}_0} z)} \right) \right\rangle_{z, \epsilon}.
\]
(13)

Here \( D x = dx e^{-x^2/2\pi} \) is a gaussian integral. Noting the factorization over \( n \) replicas, this can be simplified as
\[
e^{-E(\hat{Q})} = \left\langle \left( e^{-n \Lambda(z, \epsilon)} \right) \right\rangle_{z, \epsilon},
\]
(14)

where
\[
\Lambda(z, \epsilon) = -\log \left( \int D x e^{-\beta \rho(\epsilon + \sqrt{\Delta} x + \sqrt{\hat{Q}_0} z)} \right).
\]
(15)

Similarly we can factorize
\[
e^{S(\hat{Q})} = \int d\eta(E) \prod_{ab} du^a \prod_{a} \delta(u^a \cdot u^b - PQ_{ab}) e^{-\beta \sum_{a=1}^{n} \rho(\epsilon + \sqrt{\hat{Q}_0} z + \sqrt{\Delta} x^a) + \hat{\Delta} \sum_{a} (u^a)^2 + \hat{\Delta} \sum_{a} (u^a)^2},
\]
(16)

where \( \hat{\Delta} = \hat{Q}_1 - \hat{Q}_0 \). To factorize the above expression, we exploit the identity \( \langle (e^\gamma)^n \rangle_\eta = e^{\frac{\gamma^2}{2}} \) for unit gaussian \( \eta \). Letting \( \gamma = \sqrt{2\hat{Q}_0 \sum_{a} u^a} \)
\[
e^{S(\hat{Q})} = \left\langle \left( \int d\eta(E) \prod_{a} du^a e^{\sum_{a} \left[ \eta \sqrt{2\hat{Q}_0 u^a + \hat{\Delta} (u^a)^2 + \beta \sigma(s_j^0 - u_j^a)} \right]} \right) \right\rangle_\eta.
\]
(17)

It follows that,
\[ e^{S(\hat{Q})} = \left\langle e^{n \chi(n, s^0)} \right\rangle_{\eta, s^0}, \]

\[ \chi(\eta, s^0) = \log \left( \int d\epsilon e^{\epsilon^2/2} e^{-\beta \chi(\eta, s^0)} \right). \]

To determine \( F \), we calculate the low \( n \) Taylor expansion of \( \left\langle Z^n \right\rangle \), to do this we compute:

\[ \lim_{n \to 0} \frac{\partial E}{\partial n} = \left\langle \Lambda(z, \epsilon) \right\rangle_{z, \epsilon}, \tag{20} \]

\[ \lim_{n \to 0} \frac{\partial S}{\partial n} = \left\langle \chi(\eta, s^0) \right\rangle_{\eta, s^0}, \tag{21} \]

\[ \lim_{n \to 0} \frac{\partial (Q_{ab} Q_{ab})}{\partial n} = Q_1 \hat{Q}_1 - Q_0 \hat{Q}_0 = Q_0 \Delta + \hat{Q}_0 \Delta + \Delta \Delta. \tag{22} \]

Using these limits to compute the free energy density given by the saddle point in (10)

\[ \hat{F} = -\frac{\beta F}{N} = \min_{Q_0, Q_0, \Delta, \Delta} \left[ \left\langle \Lambda(z, \epsilon) \right\rangle_{z, \epsilon} + \frac{1}{\alpha} \left\langle Q_0 \Delta + \hat{Q}_0 \Delta + \Delta \Delta \right\rangle - \frac{1}{\alpha} \left\langle \chi(\eta, s^0) \right\rangle_{\eta, s^0} \right]. \tag{23} \]

Minimizing with respect to each of the order parameters \( Q_0, \hat{Q}_0, \Delta, \Delta \), we find the following set of coupled equations defining the order parameters:

\[ \hat{\Delta} = -\alpha \frac{\partial \left\langle \Lambda \right\rangle_{z, \epsilon}}{\partial Q_0}, \tag{24} \]

\[ \Delta = \frac{\partial \left\langle \chi \right\rangle_{\eta, s^0}}{\partial Q_0}, \tag{26} \]

\[ \hat{Q}_0 + \Delta = -\alpha \frac{\partial \left\langle \Lambda \right\rangle_{z, \epsilon}}{\partial \Delta}, \tag{25} \]

\[ Q_0 + \Delta = \frac{\partial \left\langle \chi \right\rangle_{\eta, s^0}}{\partial \Delta}. \tag{27} \]

By making a change of variables, we can write

\[ \Lambda(z, \epsilon) = -\log \left( \frac{1}{\sqrt{\Delta}} \int d\epsilon e^{-\frac{1}{\alpha \Delta} (\epsilon - \sqrt{\Delta} z - \epsilon)^2 - \beta \rho(y)} \right), \tag{28} \]

so that

\[ \frac{\partial \left\langle \Lambda \right\rangle_{z, \epsilon}}{\partial Q_0} = -\frac{1}{2\Delta\sqrt{\Delta}} \left\langle \left\langle \frac{1}{2} \int dy e^{-\frac{1}{\alpha \Delta} (\epsilon - \sqrt{\Delta} z - \epsilon)^2 - \beta \rho(y)} \right\rangle_{z, \epsilon} \right\rangle_{z, \epsilon}. \tag{29} \]

Using the fact that for a gaussian random variable \( z \), \( \left\langle z f(z) \right\rangle_z = \left\langle f'(z) \right\rangle_z \), we can re-write this as

\[ \frac{\partial \left\langle \Lambda \right\rangle_{z, \epsilon}}{\partial Q_0} = -\frac{1}{2\Delta^2} \left\langle \left\langle \frac{1}{2} \int dy e^{-\frac{1}{\alpha \Delta} (\epsilon - \sqrt{\Delta} z - \epsilon)^2 - \beta \rho(y)} \right\rangle_{z, \epsilon} \right\rangle_{z, \epsilon} - \Delta. \tag{30} \]

Defining \( \epsilon_{Q_0} = \epsilon + \sqrt{\Delta} z \) yields

\[ \hat{\Delta} = \frac{\alpha}{2\Delta^2} \left\langle \left\langle (\epsilon - \epsilon_{Q_0})^2 \right\rangle_{P_{Q_0}(y)} - \left\langle (\epsilon - \epsilon_{Q_0})^2 \right\rangle_{P_\epsilon(y)} \right\rangle_{\epsilon_{Q_0}} - \frac{\alpha}{2\Delta}, \tag{31} \]

where \( \left\langle . \right\rangle_{P_{Q_0}(y)} \) denotes average over a Gibbs distributions \( P_{Q_0}(y) \) \( \propto e^{-\beta_0 (\epsilon - \sqrt{\Delta} z - \epsilon)^2 - \beta \rho(y)} \). Similarly, we find

\[ \frac{\partial \left\langle \Lambda \right\rangle_{z, \epsilon}}{\partial \Delta} = -\frac{1}{2\Delta^2} \left\langle \left\langle \frac{1}{2} \int dy e^{-\frac{1}{\alpha \Delta} (\epsilon - \sqrt{\Delta} z - \epsilon)^2 - \beta \rho(y)} \right\rangle_{z, \epsilon} \right\rangle_{z, \epsilon} + \frac{1}{\Delta^2}, \tag{32} \]

implying

\[ \hat{Q}_0 + \Delta = \frac{\alpha}{2\Delta^2} \left\langle \left\langle (\epsilon - \epsilon_{Q_0})^2 \right\rangle_{P_{Q_0}(y)} \right\rangle_{\epsilon_{Q_0}} - \frac{\alpha}{2\Delta}. \tag{33} \]

We can split this expectation into a variance term and bias term to yield

\[ \hat{Q}_0 = \frac{\alpha}{2\Delta^2} \left\langle \left\langle (\epsilon - \epsilon_{Q_0})^2 \right\rangle_{P_{Q_0}(y)} \right\rangle_{\epsilon_{Q_0}}. \tag{34} \]

We can also compute order parameters \( (\Delta, Q_0) \), and find
\[ \Delta = \left\langle \frac{f \, du}{\sqrt{2\pi} \sigma(\eta, u)} \right\rangle, \]

where by using the property of gaussian integrals to take the derivative with respect to \( \eta \), we find

\[ \Delta = \left\langle \frac{\left( u^2 \right) P_\sigma(u)}{P_\sigma(u)} \right\rangle, \]

where \( P_\sigma(u) \propto e^{-\frac{1}{2} u^2} \). Similarly, minimizing with respect to \( \hat{\Delta} \), we find

\[ Q_0 + \Delta = \left\langle \frac{f \, due^2 \sqrt{2\pi} \sigma u^2 - \beta \sigma(y)}{f \, due^2 \sqrt{2\pi} \sigma u^2 - \beta \sigma(y)} \right\rangle \]

splitting this into a bias and variance component yields

\[ Q_0 = \left\langle \frac{\left( u^2 \right) P_\sigma(u)}{P_\sigma(u)} \right\rangle. \]

These relations, which hold for arbitrary temperature, may be written as

\[ Q_0 = \left\langle \frac{\left( u^2 \right) P_\sigma(u)}{P_\sigma(u)} \right\rangle, \quad Q_0 = \left\langle \frac{\left( u^2 \right) P_\sigma(u)}{P_\sigma(u)} \right\rangle, \]

where \( q_d = \frac{Q_0}{2\Delta^2} \), \( \delta u = u - \langle u \rangle \), \( \delta y = y - \langle y \rangle \), and

\[ P_\rho(y) \propto e^{-\frac{1}{2} (y - \langle y \rangle)^2} - \beta \rho(y), \quad P_\sigma(u) \propto e^{\frac{1}{2} (u - \langle u \rangle)^2 - \beta \sigma(u)}. \]

We make the simplifying change of variables \( \Delta \to \Lambda_\rho, \hat{\Delta} \to \frac{1}{2\Lambda_\rho}, Q_0 \to q_s \), and \( Q_0 \to \frac{q_d}{2\Delta^2} \). We also relabel our gaussian unit noise \( z \to z, \eta \to z_s \) to match the main paper.

### 2.2 Replica equations in the low temperature limit

**Result 2. The low temperature limit:**

In the low temperature limit \( \beta \to \infty \), the energy minimization problem and M-estimation problem are equivalent. We seek solutions to the replica equations where the order parameters scale with \( \beta \) as \( \Lambda_\rho = \frac{\Lambda}{\beta}, \Lambda_\sigma = \frac{\Lambda}{\beta} \), so that \( \lambda_\sigma, \lambda_\rho, q_s, q_d \) are \( O(1) \) scalars and we find that the probability distributions for \( d \) and \( s \) simplify in the low temperature limit to:

\[ P_\rho(d|\epsilon_q) = N\left( \epsilon, \frac{\lambda_\rho d^2(\epsilon_q)}{\beta} \right), \quad P_\sigma(s|s_{qd}) = N\left( \frac{\lambda_\sigma d s^2(s_{qd})}{\beta} \right), \]

\[ \hat{\epsilon}(\epsilon_q) = \mathcal{P}_{\lambda_\rho}[\rho](\epsilon_q), \quad \hat{s}(s_{qd}) = \mathcal{P}_{\lambda_\rho}[\sigma](s^0_{qd}). \]

where \( \mathcal{P}_{\lambda}[f] \) is the proximal map function (see Appendix C for more information) defined as

\[ \mathcal{P}_{\lambda}[f](x) = \arg \min_y \left[ \frac{(x - y)^2}{2\lambda} + f(y) \right], \]

and \( \epsilon_q = \epsilon + \sqrt{q_s z_s, s_{qd}^0 = s + \sqrt{q_d z_s} \) are random variables corresponding to the corrupted versions of the noise and signal where \( z_s, z_s \) defined as zero mean, unit variance gaussian noise.

**Derivation—.**

\[ P_\rho(d|\epsilon_q) \propto e^{-\beta \left( \frac{(d - \epsilon_q)^2}{2\lambda_\rho} \right) + \rho(d)} \]

\[ P_\sigma(s|s_{qd}) \propto e^{-\beta \left( \frac{(s - s_{qd})^2}{2\lambda_\rho} + \sigma(s) \right)} \]

If we define \( \hat{\epsilon} \) and \( \hat{s} \) to be the mode (arg max) of distributions \( P_\rho \) and \( P_\sigma \) respectively, which is equivalent to maximizing the argument of the exponentials, then

\[ \hat{\epsilon} = \mathcal{P}_{\lambda_\rho}[\rho](\epsilon_q), \quad \hat{s} = \mathcal{P}_{\lambda_\rho}[\sigma](s^0_{qd}). \]

Taylor expanding \( P_\rho \) and \( P_\sigma \) about these maxima yields
Differentiating with respect to \( \hat{\epsilon}_{q_0} \) the proximal point (LHS of (48)), we necessarily get zero. Applying this property to (51) yields:

\[
\frac{\hat{\epsilon} - \epsilon_{q_0}}{\lambda_{\rho}} + \rho'(\hat{\epsilon}) = 0.
\]

Differentiating with respect to \( \hat{\epsilon} \), we find

\[
\frac{1}{\lambda_{\rho}} + \rho''(\hat{\epsilon}) = \frac{1}{\lambda_{\rho}} \delta'(\hat{\epsilon}_{q_0}).
\]

Substituting this form into (53) and similarly for (54), it follows that

\[
P_\rho(d|\epsilon_{q_0}) \propto e^{-\frac{\beta}{2s} \left[ \frac{1}{\lambda_{\rho}} + \rho'(\hat{\epsilon}) \right] (d-\hat{\epsilon})^2},
\]

(53)

\[
P_\sigma(s|s_{q_0}^0) \propto e^{-\frac{\beta}{2s} \left[ \frac{1}{\lambda_{\sigma}} + \rho'(\hat{\epsilon}) \right] (s-\hat{\epsilon})^2}.
\]

(54)

We can simplify the terms containing second derivatives in (53) and (54) as follows. Note that by the definition of the proximal map, it satisfies an interesting property: by differentiating the argument in the RHS of (48) and evaluating it at the proximal point (LHS of (48)), we necessarily get zero. Applying this property to (51) yields:

\[
\hat{\epsilon} - \epsilon_{q_0} + \rho'(\hat{\epsilon}) = 0.
\]

(55)

\[
\frac{\hat{\epsilon} - \epsilon_{q_0}}{\lambda_{\rho}} + \rho'(\hat{\epsilon}) = 0.
\]

(56)

\[
P_\rho(d|\epsilon_{q_0}) \propto e^{-\frac{\beta}{2s} \left( \frac{1}{\lambda_{\rho}} + \rho'(\hat{\epsilon}) \right) (d-\hat{\epsilon})^2},
\]

(57)

\[
P_\sigma(s|s_{q_0}^0) \propto e^{-\frac{\beta}{2s} \left( \frac{1}{\lambda_{\sigma}} + \rho'(\hat{\epsilon}) \right) (s-\hat{\epsilon})^2}.
\]

(58)

Result 3. M-estimator error predictions In the regularized case, consider any convex M-estimator defined by a loss function, regularizer pair \((\rho, \sigma)\) where the parameters \((\hat{s}_1^0, ..., S_P^0)\) and noise \((\epsilon_1, ..., \epsilon_N)\) are iid random variables. The associated zero temperature replica symmetric equations simplify to:

\[
\frac{q_d}{\lambda_{\sigma}^2} = \alpha \frac{\lambda_{\rho}}{\lambda_{\rho}^2} \left\langle \left( \hat{\epsilon} - \epsilon_{q_0} \right)^2 \right\rangle, \quad q_s = \left\langle \left( \hat{s} - s^0 \right)^2 \right\rangle, \quad \lambda_p = \lambda_{\sigma} \left\langle \hat{s} \sigma(s_{q_0}^0) \right\rangle, \quad \lambda_{\rho} = \lambda_{\rho} \left\langle \hat{s} \rho(\epsilon_{q_0}) \right\rangle.
\]

(59)

(60)

(61)

where

\[
\hat{\epsilon}(\epsilon_{q_0}) = \mathcal{P}_{\lambda_{\rho}}[\rho](\epsilon_{q_0}), \quad \hat{s}(s_{q_0}^0) = \mathcal{P}_{\lambda_{\sigma}}[\sigma](s_{q_0}^0).
\]

(62)

(63)

(64)

Derivation. — This result follows from the substituting the low temperature form of \( P_\rho \) and \( P_\sigma \) in Result 2 into the finite temperature consistency equations in Result 1 and taking the zero temperature limit.

We note that these replica equations have a well-defined relationship to the original M-estimation problem in (2). Consider for example pairs of estimated and true signal components: \((\hat{s}_i^1, s_i^0)\) for \( i = 1, \ldots, P \) in the original high-dimensional inference problem. We can obtain an empirical histogram of such joint pairs, that converges as \( P \to \infty \) to a joint distribution \( P(\hat{s}_i, s_i^0) \). This distribution is self-averaging - i.e. does not depend on the detailed realization of signal, measurements, or noise. The replica theory makes a mean-field prediction for this self-averaging joint distribution through (64), as described in the main paper. This joint distribution arises as an auxiliary scalar inference problem in which gaussian noise of variance \( q_d \) is added to \( s_i^0 \), and this noise is cleaned up by a proximal descent step to obtain an estimate \( \hat{s}_i \). Integrating out the noise yields the mean field prediction \( P_{MF}(\hat{s}_i, s_i^0) \).

A similar statement holds from the perspective of noise as opposed to signal estimation. Any inference procedure that infers an estimated signal \( \hat{s} \) also implicitly estimates the noise through \( \hat{\epsilon} = y - X \hat{s} \), thus yielding a decomposition of the measurement outcome \( y \) into a sum of estimated signal contribution \( X \hat{s} \) and estimated noise contribution \( \hat{\epsilon} \). Just as we did for the estimated signal and the true signal, we could also consider the joint distribution of components of the estimated noise and true noise, \( (\epsilon_{\mu}, \epsilon_{\mu}) \) for \( \mu = 1, \ldots, N \). Again in the \( N \to \infty \) limit the empirical joint histogram of these quantities converges to a self-averaging joint distribution \( P(\epsilon_{\mu}, \epsilon_{\mu}) \). The replica theory makes a mean field prediction for this joint distribution through (63), also described in the main paper. This joint distribution arises as an auxiliary scalar inference problem in which gaussian noise of variance \( q_{\epsilon} \) is added to the true noise \( \epsilon \), and this additional noise is cleaned up by a proximal descent step to obtain an estimated noise \( \hat{\epsilon} \). Integrating out the Gaussian noise yields the mean field prediction \( P_{MF}(\hat{\epsilon}_{\mu}, \epsilon_{\mu}) \). These distributional predictions can be derived by following a standard replica method in which the above steps are modified by the introduction of a delta function picking off one component of either the estimated signal or noise. See [1] section 8.2 for an example of this modification.
Result 4. The special case of unregularized M-estimation, for which $\sigma = 0$, follows directly from Result 3. In this case:

$$\alpha \left\langle \left( \hat{\epsilon} - \epsilon_{q_s} \right)^2 \right\rangle = q_s \quad \text{and} \quad \alpha \left\langle \left( 1 - \epsilon' (\epsilon_{q_s}) \right) \right\rangle = 1 \quad \text{where}$$

$$\hat{\epsilon} (\epsilon_{q_s}) = \mathcal{P}_{\lambda_\rho} [\rho] (\epsilon_{q_s}) \quad \text{(67)}$$

Using the relation between the Moreau envelope and proximal map, we can also write these equations in the form:

$$\alpha \left\langle \left( \lambda_\rho \mathcal{M}_{\lambda_\rho} [\rho]' (\epsilon_{q_s}) \right)^2 \right\rangle = q_s \quad \text{and} \quad \alpha \left\langle \left( \lambda_\rho \mathcal{M}_{\lambda_\rho} [\rho]'' (\epsilon_{q_s}) \right) \right\rangle = 1 \quad \text{where}$$

$$\text{the mean square M-estimator error } q_s \text{ as well as the order parameter } \lambda_\rho \text{ are may be computed as the solution to this pair of equations, this result is in agreement with the findings of previous work [2].}$$

Derivation of Result 4.

The choice $\sigma = 0$ yields a linear map $\mathcal{P}_{\lambda_\rho} [\sigma] (x) = x$. Substituting this into (62) yields $\lambda_\rho = \lambda_\sigma$, and into (61) yields $q_s = q_d$. Using this simplification equations (59) and (60) become (65) and (66). Then, direct substitution of the relation between the proximal map and the Moreau envelope (74) derived in Appendix C.1 yields (68) and (69). Moreover, dividing (68) by the square of (69) yields the form of the replica equations exhibited in the main paper.

2.3 Moreau envelope formulation of the replica equations

While the above replica symmetric equations (59-61) provide a nice characterization of the mean field distributions $\mathcal{P}_{\text{MF}} (\hat{\epsilon}, \epsilon)$ and $\mathcal{P}_{\text{MF}} (\hat{s}, s^0)$, it will also be useful to formulate these replica symmetric equations in terms of Moreau envelopes, instead of proximal maps. This reformulation will simplify derivations of bounds on inference error in section 4.2 and derivations of optimal inference procedures in section 5.2. To achieve this reformulation, we systematically apply the relation between proximal maps and Moreau envelopes (74) derived in Appendix C.1. This yields the result:

Result 5. M-estimator Formulation

$$\frac{q_d}{\lambda_\sigma^2} = \alpha \left\langle \left( \mathcal{M}_{\lambda_\rho} [\rho]' (\epsilon_{q_s}) \right)^2 \right\rangle \quad \text{(70)} \quad q_s + q_d \left( 1 - 2 \frac{\lambda_\sigma}{\lambda_\rho} \right) = \lambda_\sigma^2 \left\langle \left( \mathcal{M}_{\lambda_\rho} [\sigma]' (s_{q_d}^0) \right)^2 \right\rangle \quad \text{(72)}$$

$$\frac{1}{\lambda_\sigma} = \alpha \left\langle \left( \mathcal{M}_{\lambda_\rho} [\rho]'' (\epsilon_{q_s}) \right) \right\rangle \quad \text{(71)} \quad \frac{\lambda_\rho}{\lambda_\sigma} - 1 = \lambda_\sigma \left\langle \left( \mathcal{M}_{\lambda_\rho} [\sigma]'' (s_{q_d}^0) \right) \right\rangle \quad \text{(73)}$$

Derivation.—

In Appendix C.1 we derive the relation between the proximal map and Moreau envelope:

$$\mathcal{P}_{\lambda} [f] (x) = x - \lambda \mathcal{M}_{\lambda} [f]' (x). \quad \text{(74)}$$

By applying the above relation to (59) and (60) respectively we arrive at (70) and (71). Next we derive (72), starting with (62):

$$\lambda_\rho = \lambda_\sigma \left\langle \left( \mathcal{P}_{\lambda_\rho} [\sigma]' (s_{q_d}^0) \right) \right\rangle = \lambda_\sigma \left( 1 - \left\langle \left( \lambda_\sigma \mathcal{M}_{\lambda_\rho} [\sigma]' (s_{q_d}^0) \right) \right\rangle \right).$$

Rearranging yields

$$1 - \frac{\lambda_\rho}{\lambda_\sigma} = \lambda_\sigma \left\langle \left( \mathcal{M}_{\lambda_\rho} [\sigma]'' (s_{q_d}^0) \right) \right\rangle. \quad \text{(76)}$$

Finally, to derive (73), we begin with (61) to obtain

$$q_s = \left\langle \left( \mathcal{P}_{\lambda_\rho} [\sigma] (s^0 + \sqrt{q_d} z) - s^0 \right)^2 \right\rangle_{s^0, z} = \left\langle \left( \sqrt{q_d} z - \lambda_\sigma \mathcal{M}_{\lambda_\rho} [\sigma]' (s^0 + \sqrt{q_d} z) \right)^2 \right\rangle_{s^0, z}. \quad \text{(77)}$$

By expanding the square above we find

$$q_s = q_d - 2 \sqrt{q_d} \left\langle \left( z \lambda_\sigma \mathcal{M}_{\lambda_\rho} [\sigma]' (s^0 + \sqrt{q_d} z) \right) \right\rangle_{s^0, z} + \lambda_\sigma^2 \left\langle \left( \mathcal{M}_{\lambda_\rho} [\sigma]' (s_{q_d}^0) \right)^2 \right\rangle_{s_{q_d}^0}. \quad \text{(78)}$$

Now, we again use the fact that for a gaussian random variable $z$, $\left\langle (zf(z)) \right\rangle_z = \left\langle f'(z) \right\rangle_z$, to show that
\[
\sqrt{q_d} \langle \langle z \lambda_\sigma M_{\lambda_\sigma} [\sigma'] (s^0 + \sqrt{q_d} z) \rangle \rangle \rangle_z = q_d \langle \langle \lambda_\sigma M_{\lambda_\sigma} [\sigma''(s_{q_d}^0)] \rangle \rangle_z = q_d \left(1 - \frac{\lambda_\sigma}{\lambda_\rho}\right),
\]
where the last equality follows from (73). Substituting this into (78) yields
\[
q_s = q_d - 2 q_d \left(1 - \frac{\lambda_\sigma}{\lambda_\rho}\right) + \lambda_\sigma^2 \langle \langle (M_{\lambda_\sigma} [\sigma'](s_{q_d}^0)) 2 \rangle \rangle_{s_{q_d}^0}.
\]
which may be rearranged as
\[
q_s + q_d \left(1 - 2 \frac{\lambda_\sigma}{\lambda_\rho}\right) = \lambda_\sigma^2 \langle \langle (M_{\lambda_\sigma} [\sigma'](s_{q_d}^0)) 2 \rangle \rangle_{s_{q_d}^0}.
\]
Finally, note that dividing (70) by the square of (71) yields the form of the replica equation for \(q_d\) exhibited in the main paper.

### 2.4 Accuracy of inference from the perspective of noise

Just as \(q_s\) measures inference performance through the MSE (Mean Squared Error) in estimating the signal, it is also useful to capture inference performance through the MSE in estimating the noise, which we work out in this section. This measure of inference performance, from the perspective of noise, will yield insights into the nature of optimal inference, as well as the nature of overfitting and generalization (see sections 5.1 and 5.2 below).

**Result 6.** We define \(q_e = \langle\langle (\epsilon(e_{q_s}) - e)^2 \rangle\rangle\), which corresponds to replica prediction for the MSE of an M-estimator in estimating the the noise \(e\). In the regularized case we find:
\[
q_e = q_s + q_s \alpha \left[ q_d \left( \frac{\lambda_\sigma}{\lambda_\rho} - \frac{q_s}{q_d} \right)^2 - \frac{q_s}{q_d} \right].
\]

In the unregularized case and the over-sampled regime \((\alpha > 1)\), \(q_d = q_s\) and \(\lambda_\rho = \lambda_\sigma\), so this reduces to
\[
q_e = q_s \left( \frac{\alpha - 1}{\alpha} \right).
\]

*Derivation.*—

Using the definition of \(q_e\) and the relationship between the Moreau envelope and proximal map (74), we obtain
\[
q_e = \langle\langle (\epsilon(e_{q_s}) - e)^2 \rangle\rangle = \langle\langle (\epsilon(e_{q_s}) + \sqrt{q_s} z)^2 \rangle\rangle = \langle\langle (\sqrt{q_s} z - \lambda_\rho M_{\lambda_\rho} [\rho] (e_{q_s}))^2 \rangle\rangle.
\]
Expanding the square above yields
\[
q_e = q_s - 2 \sqrt{q_s} \lambda_\rho \langle\langle z M_{\lambda_\rho} [\rho] (e_{q_s}) \rangle\rangle + \lambda_\rho^2 \langle\langle (M_{\lambda_\rho} [\rho] (e_{q_s}))^2 \rangle\rangle.
\]

Applying the identity \(\langle z f(z) \rangle_z = \langle f'(z) \rangle_z\) where \(z\) is a unit gaussian random variable yields
\[
q_e = q_s - 2 q_s \langle\langle \lambda_\rho M_{\lambda_\rho} [\rho] (e_{q_s}) \rangle\rangle + \langle\langle (\lambda_\rho M_{\lambda_\rho} [\rho] (e_{q_s}))^2 \rangle\rangle.
\]

Then, substituting the replica symmetry equations (70),(71), yields
\[
q_e = q_s - q_s \frac{q_s}{\alpha \lambda_\rho} + q_d \frac{\lambda_\rho^2}{\alpha \lambda_\sigma^2} = q_s + q_s \alpha \left[ \frac{q_d \left( \frac{\lambda_\rho}{\lambda_\sigma} - \frac{q_s}{q_d} \right)^2 - \frac{q_s}{q_d} \right].
\]

### 2.5 Noise-free replica equations

Motivated by the theory of compressed sensing which yields the interesting result that we can sometimes exactly recover unknown sparse signals in the undersampled measurement regime \(\alpha < 1\) in the absence of noise via \(L_1\) minimization, we consider the more general case of arbitrary convex inference in the absence of noise.
Result 7. In the limit of noise-free measurements, $y_\mu = X_\mu \cdot s^0$. The following equalities hold in the regime where the MSE error is non-zero

$$\hat{s} = \min_s \sum_{j=1}^P \sigma(s_j) \quad s.t \quad y = Xs.$$  \hspace{1cm} (88)

$$q_d = \frac{1}{\alpha} q_\sigma,$$  \hspace{1cm} (89)

$$q_s = \langle\langle (\hat{s}(s^0_{q_d}) - s^0)^2 \rangle\rangle_{s^0, z},$$  \hspace{1cm} (91)

$$\lambda_\mu = \alpha \lambda_\sigma,$$  \hspace{1cm} (90)

$$\langle\langle \hat{s}'(s^0_{q_d}) \rangle\rangle_{s^0_{q_d}} = \alpha,$$  \hspace{1cm} (92)

where $\hat{s}(s^0_{q_d}) = P_{\lambda_\sigma}[^{\sigma}_{\lambda_\sigma}](s^0_{q_d}).$  \hspace{1cm} (93)

Derivation.—

The problem is equivalent to a M-estimator of the form

$$\hat{s} = \arg\min_s \left[ \sum_{\mu} \gamma \left( y_\mu - X_{\mu j} s_j \right)^2 + \sum_j \sigma(s_j) \right],$$  \hspace{1cm} (94)

where we will take the limit $\gamma \to \infty$ to strictly enforce equality on the constraint equations. Before we take this limit, we substitute the form of $\rho$ into Result 3, yielding the following 4 equations:

$$\frac{q_d}{\lambda_\sigma^2} = \frac{\alpha q_s \gamma^2}{(1 + \gamma \lambda_\rho)^2}$$  \hspace{1cm} (95)

$$q_s = \langle\langle \left( P_{\lambda_\sigma}[^{\sigma}_{\lambda_\sigma}](s^0_{q_d}) - s^0 \right)^2 \rangle\rangle_{s^0, z},$$  \hspace{1cm} (97)

$$\frac{1}{\lambda_\sigma} = \frac{\alpha \gamma}{1 + \gamma \lambda_\rho},$$  \hspace{1cm} (96)

$$\lambda_\rho = \lambda_\sigma \langle\langle \left( P_{\lambda_\sigma}[^{\sigma}_{\lambda_\sigma}](s^0_{q_d}) \right)' \rangle\rangle_{s^0_{q_d}}.$$  \hspace{1cm} (98)

One can substitute (96) into (95) to show

$$q_s = \alpha q_d.$$  \hspace{1cm} (99)

Now we consider two cases as $\gamma \to \infty$. First, in the finite error regime, $q_s$ remains $O(1)$ as $\gamma \to \infty$, $\gamma \lambda_\rho \to \infty$. Second, in the error-free regime $q_s \to 0$, while $\gamma \lambda_\rho$ remains $O(1)$. In the finite error regime, the equations simplify and (96) becomes $\lambda_\rho = \alpha \lambda_\sigma$. Then (98) becomes

$$\langle\langle \left( P_{\lambda_\sigma}[^{\sigma}_{\lambda_\sigma}](s^0_{q_d}) \right)' \rangle\rangle_{s^0_{q_d}} = \alpha.$$  \hspace{1cm} (100)

Thus we obtain all the results in the box above corresponding to the finite error regime. We note that in the zero error limit, (96) becomes

$$\lambda_\sigma = \frac{1 + \gamma \lambda_\rho}{\alpha \gamma} \to 0,$$  \hspace{1cm} (101)

since $\gamma \lambda_\rho = O(1)$ as $\gamma \to \infty$. Applying the fact that $\lambda_\sigma = 0$ to (97) yields $q_s = q_d$, and to (98) yields $\lambda_\rho = \lambda_\sigma = 0$. Thus we find that in the zero error regime, as expected $\lambda_\rho, \lambda_\sigma, q_d, q_s \to 0$ is a solution to the RS equations. See [3] for a more detailed study of noise-free perfect inference for the specific case of $L_1$ minimization.
3 Analytic example: quadratic loss and regularization

3.1 General quadratic MSE

Result 8. As an example, we derive the asymptotic MSE, \( q_s \), analytically for the case of M-estimators with a square loss function and quadratic regularization:

\[
\rho(x) = \frac{x^2}{2}, \quad \sigma(x) = \frac{x^2}{2},
\]

we find the normalized MSE (i.e. the fraction of unexplained variance) to be

\[
\bar{q}_s = \frac{q_s}{\langle \langle (s^0)^2 \rangle \rangle} = \frac{\phi + \alpha \gamma^2 \lambda^2}{\alpha (1 + \gamma \lambda \sigma)^2 - 1}
\]

where \( \phi = \frac{\langle \langle (\epsilon^2) \rangle \rangle}{\langle \langle (s^0)^2 \rangle \rangle} = \frac{1}{SNR} \) is the inverse signal to noise ratio, and \( \lambda \) is the non-negative root which solves the quadratic equation:

\[
\alpha \gamma \lambda^2 + (\alpha - \gamma - 1) \lambda - 1 = 0.
\]

Derivation. —

From Result 3, the fact that \( f(x) = \frac{3}{2}x^2 \), \( \mathcal{P}_\lambda[f](x) = \frac{x^2}{1 + \alpha \gamma} \), allows us to rewrite (63) and (64) as

\[
\hat{\epsilon}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_p}[\rho](\epsilon_{q_s}) = \frac{\epsilon_{q_s}}{1 + \lambda_p}, \quad \hat{s}(s_{q_s}) = \mathcal{P}_{\lambda_p}[\sigma](s_{q_s}) = \frac{s_{q_s}}{1 + \gamma \lambda \sigma}. \tag{106}
\]

Substituting the form of \( \hat{\epsilon}, \hat{s} \) into (59 - 62) yields

\[
\frac{q_d}{\lambda^2} = \frac{\alpha}{\alpha} = \frac{\alpha}{(1 + \lambda_p \sigma)^2} \langle \langle \epsilon_{q_s}^2 \rangle \rangle_{\epsilon_{q_s}}, \tag{107}
\]

\[
\frac{1}{\lambda} = \frac{\alpha}{1 + \lambda_p} \implies \lambda = \frac{1 + \lambda_p}{\alpha} \tag{108}
\]

which implies

\[
\frac{q_d}{\lambda^2} = \frac{\langle \langle \epsilon_{q_s}^2 \rangle \rangle}{\alpha} = \frac{\langle \langle \epsilon^2 \rangle \rangle}{\alpha} + q_s \tag{109}
\]

\[
q_s = \frac{q_d + (\lambda \gamma)^2 \langle \langle (s^0)^2 \rangle \rangle_{s^0}}{(1 + \lambda \gamma)^2} \tag{110}
\]

Combining (108) and (111) we find that \( \lambda \) must satisfy the quadratic equation

\[
\alpha \gamma \lambda^2 + (\alpha - \gamma - 1) \lambda - 1 = 0. \tag{113}
\]

The constraint of non-negativity forces \( \lambda \) to be the positive root, which is unique. Finally, substituting (109) into (112) yields

\[
q_s = \frac{\langle \langle \epsilon^2 \rangle \rangle + q_s + \alpha (\gamma \lambda \sigma)^2 \langle \langle (s^0)^2 \rangle \rangle}{\alpha (1 + \gamma \lambda \sigma)^2} \tag{114}
\]

Re-arranging to solve for \( q_s \) and dividing by the variance of the parameters, we find

\[
\bar{q}_s = \frac{q_s}{\langle \langle (s^0)^2 \rangle \rangle} = \frac{\phi + \alpha (\gamma \lambda \sigma)^2}{\alpha (1 + \gamma \lambda \sigma)^2 - 1}, \tag{115}
\]

where we define \( \phi = \frac{\langle \langle (\epsilon^2) \rangle \rangle}{\langle \langle (s^0)^2 \rangle \rangle} \) to be the noise to signal ratio. Note that in the limit of strong regularization (\( \gamma \to \infty \)), this reduces to \( q_s = \frac{\langle \langle (s^0)^2 \rangle \rangle}{\alpha} \) which is intuitive since the strong ridge penalty pins the estimator \( \hat{s} \) to 0. The other extreme is the un-regularized M-estimator, which corresponds simply to least squares regression and yields

\[
q_s = \frac{\langle \langle \epsilon^2 \rangle \rangle}{\alpha - 1}, \tag{116}
\]

which demonstrates that the MSE is unbounded as the number of samples approaches the number of dimensions.
3.2 Optimal quadratic MSE

Result 9. **Optimal Quadratic M-estimation:** We find that the optimal quadratic M-estimator has the form:

\[
\rho(x) = \frac{x^2}{2}, \quad \sigma(x) = \phi \frac{x^2}{2},
\]

and satisfies

\[
\bar{q}_s = 1 - \alpha - \phi + \frac{(\phi + \alpha - 1)^2 + 4\phi}{2}.
\]

This result does not depend on the general form of \( P_e \) and \( P_s \) except via \( \phi \), the inverse SNR.

**Derivation of Optimal Regularization Parameter:**

To find the value of \( \gamma \) that minimizes \( \bar{q}_s \), we set the derivative of (103) with respect to \( \gamma \) equal to zero:

\[
\frac{d\bar{q}_s}{d\gamma} = \frac{2\alpha(\lambda_\sigma + \gamma \frac{\lambda_\sigma}{\phi})}{(\alpha(1 + \gamma \lambda_\sigma)^2 - 1)^2} (\alpha \gamma^2 \lambda_\sigma^2 + (\alpha - \phi - 1)\gamma \lambda_\sigma - \phi) = 0.
\]

We look for when

\[
\alpha \gamma^2 \lambda_\sigma^2 + (\alpha - \phi - 1)\gamma \lambda_\sigma - \phi = 0,
\]

we then multiply (104) by \( \gamma \):

\[
\alpha \gamma^2 \lambda_\sigma^2 = \gamma - (\alpha - \gamma - 1)\gamma \lambda_\sigma.
\]

Substituting this result into (120) yields

\[
(\gamma - \phi)(1 + \gamma \lambda_\sigma) = 0,
\]

so we find a fixed point of \( \bar{q}_s \) occurs at \( \gamma = \phi \).

**Derivation of Optimal \( \bar{q}_s \)**

Substituting the optimal regularization parameter \( \gamma = \phi \) into (104) and multiplying by an overall factor of \( \phi \) yields

\[
\alpha(\phi \lambda_\sigma)^2 + (\alpha - \phi - 1)\phi \lambda_\sigma - \phi = 0,
\]

multiplying both sides by the positive factor \( 1 + \phi + \alpha \), yields

\[
\alpha(1 + \phi + \alpha)(\phi \lambda_\sigma)^2 + (1 + \phi + \alpha)(\alpha - \phi - 1)\phi \lambda_\sigma - \phi(\phi + \alpha + 1) = 0.
\]

This can be expanded an re-arranged as

\[
\phi(\alpha \lambda_\sigma - 1)(\phi \lambda_\sigma(1 + \phi + \alpha) + \phi + \alpha - 1) = \phi \lambda_\sigma(1 + \phi - \alpha) + 2\phi.
\]

It follows that

\[
\phi(\alpha \lambda_\sigma - 1) = \frac{\phi \lambda_\sigma(1 + \phi - \alpha) + 2\phi}{(\phi \lambda_\sigma(1 + \phi + \alpha) + \phi + \alpha - 1)}.
\]

Substitution of the the form of \( (\phi \lambda_\sigma)^2 \) from (123) into (103) yields

\[
\bar{q}_s = \frac{\phi + \alpha \phi^2 \lambda_\sigma^2}{\alpha(1 + \phi \lambda_\sigma)^2 - 1} = \frac{\phi \lambda_\sigma(1 + \phi - \alpha) + 2\phi}{(\phi \lambda_\sigma(1 + \phi + \alpha) + \phi + \alpha - 1)}.
\]

Combining the previous two equations implies that

\[
\bar{q}_s = \phi(\alpha \lambda_\sigma - 1).
\]

We complete the derivation by substituting the value of \( \phi \lambda_\sigma \), which is the positive root of a quadratic equation (123):

\[
\phi \lambda_\sigma = \frac{\phi + 1 - \alpha + \sqrt{(\phi + 1 - \alpha)^2 + 4\alpha \phi}}{2\alpha}.
\]

Substituting (129) into (128) yields
\[
\bar{q}_s = \frac{\phi + 1 - \alpha + \sqrt{(\phi + 1 - \alpha)^2 + 4\alpha \phi - 2\phi}}{2} = 1 - \alpha - \phi + \sqrt{(\phi + \alpha - 1)^2 + 4\phi}.
\] (130)

### 3.3 Training and test errors

\[
\bar{E}_{\text{test}} = \frac{\bar{E}_{\text{test}}}{\langle (s^0)^2 \rangle} = \bar{q}_s + \phi,
\] (131)

\[
\bar{E}_{\text{train}} = \langle (\mathcal{P}_s [\rho] (\epsilon_q)^2) \rangle.
\] (132)

Under our choice of \( \rho(x) = \frac{x^2}{2} \), it follows that

\[
\bar{E}_{\text{train}} = \langle \epsilon^2 \rangle + \bar{q}_s (1 + \lambda \rho)^2.
\] (133)

Combining (108) and (111) to find a quadratic equation in \( \lambda \rho \) and assuming the optimal setting \( \gamma = \phi \) yields

\[
(\phi \lambda \rho)^2 + (\phi \lambda \rho)(\alpha + \phi - 1) - \phi = 0.
\] (134)

This implies that in the optimal quadratic setting

\[
\lambda \rho = \bar{q}_s \phi,
\] (135)

solving for \( \bar{E}_{\text{train}} = \frac{\bar{E}_{\text{train}}}{\langle (s^0)^2 \rangle} \) yields

\[
\bar{E}_{\text{train}} = \frac{\phi^2}{\bar{E}_{\text{test}}}.
\] (136)

### 3.4 Limiting cases and scaling

<table>
<thead>
<tr>
<th>Result 10. Limiting case for optimal quadratic inference:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In the limit SNR ( \gg 1 ):</strong></td>
</tr>
<tr>
<td>( \bar{q}_s = \begin{cases} 1 - \alpha &amp; \alpha &lt; 1, \ 1 &amp; \alpha = 1, \ \frac{1}{\sqrt{\text{SNR}(\alpha - 1)}} &amp; \alpha &gt; 1 \end{cases} )</td>
</tr>
</tbody>
</table>

The following form will be convenient for expanding \( \bar{q}_s \) in different limits. Factoring \( \alpha + \phi - 1 \) out of (118) yields,

\[
\bar{q}_s = \left( \frac{\alpha + \phi - 1}{2} \right) \left( -1 + \frac{\alpha + \phi - 1}{|\alpha + \phi - 1|} \sqrt{1 + \frac{4\phi}{(\alpha + \phi - 1)^2}} \right).
\] (140)

**SNR \( \gg 1 \):**

First consider \( \alpha = 1 \), in which case (140) reduces to

\[
\bar{q}_s = \phi \left( -1 + \sqrt{1 + \frac{1}{\phi}} \right) = \frac{1}{2\text{SNR}} (\sqrt{1 + 4\text{SNR}} - 1).
\] (141)

In order to Taylor expand this in small \( \phi \), we write the above expression as

\[
\bar{q}_s = \frac{\phi}{2} \left( -1 + \left( \frac{4}{\phi} \right)^{1/2} \sqrt{1 + \frac{\phi}{4}} \right) \approx \frac{\phi}{2} \left( -1 + \left( \frac{1}{\phi} \right)^{1/2} \left( 1 + \frac{\phi}{8} \right) \right),
\] (142)

which approaches \( \bar{q}_s = \sqrt{\phi} = \frac{1}{\sqrt{\text{SNR}}} \) in the large SNR limit. Next, in the case \( \alpha < 1 \), we expand (140) in small \( \phi \) which yields...
\[
\bar{q}_s = \left( \frac{\alpha + \phi - 1}{2} \right) \left( -1 - \sqrt{1 + \frac{4\phi}{(\alpha + \phi - 1)^2}} \right) \approx 1 - \alpha.
\] (143)

For \(\alpha > 1\), the sign inside the square root changes, and expanding in small \(\phi\) yields
\[
\bar{q}_s = \left( \frac{\alpha + \phi - 1}{2} \right) \left( -1 + \sqrt{1 + \frac{4\phi}{(\alpha + \phi - 1)^2}} \right) \approx \frac{\phi}{\alpha + \phi - 1} \rightarrow \frac{\phi}{\alpha - 1}.
\] (144)

**Finite SNR:**
Consider the limit of finite SNR with a very low sampling rate \(\alpha \ll 1\), in which case (140) may be written as
\[
\bar{q}_s = \frac{1 - \alpha - \phi + \sqrt{\phi^2 + 2\phi - 1}}{2} = \frac{1 - \alpha - \phi + (1 + \phi)\sqrt{1 + \frac{2\phi}{(1+\phi)^2}}}{2}.
\] (145)

Expanding this expression to first order in \(\alpha\) yields
\[
\bar{q}_s = \frac{1 - \alpha - \phi + (1 + \phi)\left( 1 + \frac{\phi - 1}{1+\phi} \right)}{2} = \frac{2 - \alpha + \frac{\phi}{1 + \phi}}{2} = 1 - \frac{\alpha}{\alpha + \phi} = 1 - \frac{\mathrm{SNR}}{\alpha + \phi}.
\] (146)

These results characterize the way that, with greater SNR, increasing the measurement density yields a greater reduction in the estimation error. The case of \(\alpha = 1\) is already shown in (141). In the limit of \(\alpha \gg 1\) with finite SNR:
\[
\bar{q}_s = \left( \frac{\alpha + \phi - 1}{2} \right) \left( -1 + \sqrt{1 + \frac{4\phi}{(\alpha + \phi - 1)^2}} \right) \approx \left( \frac{\alpha + \phi - 1}{2} \right) \left( -1 + \left( 1 + 2\frac{\phi}{(\alpha + \phi - 1)^2} \right) \right) = \frac{1}{\mathrm{SNR}(\alpha - 1) + 1}.
\] (147)

**3.5 Quadratic inference through a random matrix theory perspective**
M-estimation under the choice \(\rho(x) = \frac{x^2}{2}\) and \(\sigma(x) = \gamma x^2\) becomes
\[
\hat{s} = \arg\min_s \frac{1}{2} \left[ \sum_{\mu} \left( y_{\mu} - \sum_j X_{\mu j} s_j \right)^2 + \gamma \sum_j s_j^2 \right],
\] (149)
which can be analyzed by differentiating the argument of the above expression with respect to \(s\) and setting the result equal to zero. This yields an expression for the estimated parameters:
\[
\hat{s} = \left( X^T X + \gamma I \right)^{-1} X^T y.
\] (150)

Substituting \(y = Xs^0 + \epsilon\) into the previous equation, we write this in terms of the difference between the estimated and true parameters:
\[
\hat{s} - s^0 = \left( X^T X + \gamma I \right)^{-1} \left( X^T \epsilon - \gamma s^0 \right).
\] (151)

Calculating the MSE, we find
\[
q_s = \frac{1}{P} \langle \langle \sum_i (\hat{s}_i - s^0_i)^2 \rangle \rangle = \frac{1}{P} \langle \langle \left( X^T \epsilon - \gamma s^0 \right)^T \left( X^T X + \gamma I \right)^{-T} \left( X^T X + \gamma I \right)^{-1} \left( X^T \epsilon - \gamma s^0 \right) \rangle \rangle.
\] (152)

Averaging over \(\epsilon, s^0\) yields
\[
Pq_s = \sigma^2 \text{Tr} \left( X \left( X^T X + \gamma I \right)^{-2} X^T \right) + \gamma^2 \sigma^2 \text{Tr} \left( \left( X^T X + \gamma I \right)^{-2} \right),
\] (153)
\[
= \sigma^2 \text{Tr} \left( X^T X \left( X^T X + \gamma I \right)^{-2} \right) + \gamma^2 \sigma^2 \text{Tr} \left( \left( X^T X + \gamma I \right)^{-2} \right).
\] (154)
Since $X^T X$ is a symmetric covariance matrix, it has an eigenvalue decomposition of the form

$$X^T X = Q^T A Q,$$

where $A$ is a diagonal matrix of eigenvalues. Substituting the eigenvalue decomposition of $X^T X$ into (154) yields

$$P q_s = \sigma_z^2 \text{Tr} \left[ (A + \gamma I)^{-2} \right] + \gamma^2 \sigma_z^2 \text{Tr} \left[ (A + \gamma I)^{-2} \right].$$

This allows us to write the simplified asymptotic form:

$$q_s = \sigma_z^2 \frac{1}{\alpha J} \left[ \epsilon + \sqrt{q_s} z \right] \geq \frac{1}{(\alpha - 1)J \epsilon}. \quad (155)$$

The form of the Cramer-Rao bound in this problem is derived in Appendix B.5 to be:

$$q_s \geq \frac{1}{\alpha J \epsilon}. \quad (164)$$

Thus, the Cramer-Rao bound is not saturated for finite $\alpha$. See Appendix B for more information on the Fisher information and Cramer-Rao bound.

4 Lower bounds on the error of any convex inference procedure

4.1 Unregularized inference

| Result 11. Unregularized High Dimensional Error Bound | In the over-sampled ($N > P$), but still high dimensional regime $\alpha > 1$, the asymptotic lower bound $q_s$ (estimator MSE) for any unregularized convex M-estimator is given by:

$$q_s \geq \frac{1}{\alpha J \epsilon} \geq \frac{1}{(\alpha - 1)J \epsilon}. \quad (163)$$

The form of the Cramer-Rao bound in this problem is derived in Appendix B.5 to be:

$$q_s \geq \frac{1}{\alpha J \epsilon}. \quad (164)$$

Thus, the Cramer-Rao bound is not saturated for finite $\alpha$. See Appendix B for more information on the Fisher information and Cramer-Rao bound. |

We define $f_{q_s}$ to be the pdf of $\epsilon + \sqrt{q_s} z$ where $z$ is a unit gaussian random variable. The bound above follows from constraints (65) and (66) derived in Result 4, which may be written as

$$\int f_{q_s} (y) r'(y) dy = \frac{1}{\alpha}, \quad (165) \quad \int f_{q_s} (y) r(y)^2 dy = \frac{q_s}{\alpha}. \quad (166)$$
where we have defined
\[ r(y) = \lambda_{\text{M}} \lambda_{\rho} \rho(y). \]  
(167)

Integrating (165) by parts and squaring the result yields
\[ \left(\frac{1}{\alpha}\right)^2 = \left(\int dy r(y) f_{q_s}(y) \right)^2. \]  
(168)

By breaking the integrand into two parts and applying the Cauchy-Schwartz inequality, we find:
\[ \left(\frac{1}{\alpha}\right)^2 = \left(\int dy \frac{f_{q_s}'(y)}{f_{q_s}(y)} \sqrt{f_{q_s}(y)} r(y) \right)^2 \leq J [\epsilon + \sqrt{q_s z}] \int f_{q_s}(y) r(y)^2 dy. \]  
(169)

Substituting (166) into this yields the inequality
\[ q_s \geq \frac{1}{\alpha J [\epsilon + \sqrt{q_s z}]} \]  
(170)

To complete the derivation, we apply the convolutional Fisher inequality (see Appendix B.2)
\[ J [\epsilon_1 + \epsilon_2] \leq \frac{J [\epsilon_1] J [\epsilon_2]}{J [\epsilon_1] + J [\epsilon_2]}, \]  
(171)

where \( \epsilon_1, \epsilon_2 \) are random variables drawn from different distributions. This result implies
\[ J [\epsilon + \sqrt{q_s z}] \leq \frac{J [\epsilon]}{1 + q_s J [\epsilon]}. \]  
(172)

Substituting this into (170) yields
\[ q_s \geq \frac{1}{\alpha J [\epsilon + \sqrt{q_s z}]} \geq \frac{1 + q_s J [\epsilon]}{\alpha J [\epsilon]}. \]  
(173)

This implies the inequality \( q_s J [\epsilon] \geq \frac{1}{\alpha} (1 + q_s J [\epsilon]), \) which may be applied repeatedly to the RHS of (173) to show
\[ q_s \geq \frac{1}{\alpha J [\epsilon + \sqrt{q_s z}]} \geq \frac{1}{\alpha J [\epsilon]} \left(1 + \frac{1}{\alpha} + \frac{1}{\alpha^2} + \ldots\right). \]  
(174)

For \( \alpha > 1, \) we can write this as
\[ q_s \geq \frac{1}{\alpha J [\epsilon + \sqrt{q_s z}]} \geq \frac{1}{(\alpha - 1)J [\epsilon]}. \]  
(175)

### 4.2 Regularized inference

**Result 12. High Dimensional Regularized Bounds**
For regularized M-estimators (regularizer is not assumed to be zero), we find that the estimator MSE \( q_s \) is lower bounded by
\[ q_s \geq q_d - \frac{q_d^2}{J [\epsilon + \sqrt{q_d z}]} = q_s^{\text{MSE}}(q_d), \quad \text{where} \quad q_d = \frac{1}{\alpha J [\epsilon + \sqrt{q_s z}]}. \]  
(176)

Here \( q_s^{\text{MSE}}(q_d) \) denotes the scalar minimum mean squared error of reconstructing random variable \( s^0 \) corrupted by gaussian noise with variance \( q_d \) via Bayesian inference of the posterior mean:
\[ q_s^{\text{MSE}}(q_d) = \left\langle \left\langle \left( s^0 - \langle s|s^0 + \sqrt{q_d z} \rangle \right)^2 \right\rangle \right\rangle_{s^0, z}. \]  
(177)

The above bound also leads to the inequality:
\[ q_s \geq \frac{1}{\alpha J [\epsilon + \sqrt{q_s z}] + J [s^0]}, \]  
(178)

which implies that regularized M-estimators can leverage a highly informative signal distribution to improve on the Cramer-Rao bound in the high dimensional setting.
In this result, we derive a lower bound on MSE $q_s$ over all convex Regularized M-estimators defined by pairs $(\rho, \sigma)$, given the constraints imposed by Result 5. For notational convenience, we define:

$$r_\rho(y) = \lambda_\rho \mathcal{M}_\lambda[\rho]'(y), \quad (179)$$
$$r_\sigma(y) = \lambda_\sigma \mathcal{M}_\lambda[\sigma]'(y). \quad (180)$$

We also find it convenient to define $f_{q_s}, g_{q_d}$ to be the pdf’s for $\epsilon + \sqrt{q_s} z_\epsilon$ and $s^0 + \sqrt{q_d} z_s$ respectively where $z_\epsilon, z_s$ are independent unit gaussian random variables. Thus, (72) and (73) may be written as

$$\frac{\lambda_\rho}{\lambda_\sigma} - 1 = \int g_{q_d}(y) r_\rho(y) dy, \quad (181)$$
$$\int g_{q_d}(y) r_\sigma(y)^2 dy = q_s - q_d - 2q_d \int g_{q_d}(y) r_\rho(y) dy = q_s - q_d - 2q_d \left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right). \quad (182)$$

Squaring (181) and applying the Cauchy-Schwartz inequality yields

$$\left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right)^2 = \left(\int \frac{g_{q_d}(y)}{\sqrt{q_{q_d}(y)}} \sqrt{g_{q_d}(y) r_\sigma(y) dy}\right)^2 \leq J \left[ s^0 + \sqrt{q_d} z \right] \int g_{q_d}(y) r_\sigma(y)^2 dy. \quad (183)$$

Substituting (182) into the RHS of the above inequality yields

$$\left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right)^2 \leq J \left[ s^0 + \sqrt{q_d} z \right] \left( q_s - q_d - 2q_d \left(\frac{\lambda_\rho}{\lambda_\sigma} - 1\right) \right). \quad (184)$$

Completing the square, we find

$$\left(\frac{\lambda_\rho}{\lambda_\sigma} - 1 + q_d J \left[ s^0 + \sqrt{q_d} z \right] \right)^2 \leq J \left[ s^0 + \sqrt{q_d} z \right] \left( q_s - q_d + q_d^2 J \left[ s^0 + \sqrt{q_d} z \right] \right). \quad (185)$$

Since the LHS and Fisher information are non-negative, this implies (176):

$$q_s \geq q_d - q_d^2 J \left[ s^0 + \sqrt{q_d} z \right]. \quad (186)$$

To derive the form of $q_d$ in (176), we use the two constraint equations containing $r_\rho$ (60) and (59). These may be written, by integrating (258) and (257) by parts, as

$$\int f_{q_s}(y) r_\rho(y) dy = -\frac{\lambda_\rho}{\alpha \lambda_\sigma}, \quad (187) \quad \int f_{q_s}(y) r_\rho(y)^2 dy = \frac{q_d \lambda_\rho^2}{\alpha \lambda_\sigma^2}. \quad (188)$$

Squaring (187) and applying Cauchy Schwartz as before, yields

$$\left(\frac{\lambda_\rho}{\alpha \lambda_\sigma}\right)^2 = \left(\int f_{q_s}(y) r_\rho(y) dy\right)^2 \leq J \left[ \epsilon + \sqrt{q_s} z \right] \int f_{q_s}(y) r_\rho(y)^2 dy. \quad (189)$$

By inserting (188) into the RHS, we find

$$q_d \geq \frac{1}{\alpha J \left[ \epsilon + \sqrt{q_s} z \right]}. \quad (190)$$

By applying the relation between MMSE and Fisher information in Appendix B.4 to (186), we find that $q_s$ is lower bounded by the MMSE error of reconstruction of $s^0$ given gaussian noise of variance $q_d$:

$$q_s \geq q_s^{\text{MMSE}}(q_d). \quad (191)$$

Note that the function on the RHS monotonically increasing in $q_d$ since increasing the variance in the noise will make reconstruction more difficult. It follows that the strongest bound occurs for the largest possible value of $q_d$:

$$q_d = \frac{1}{\alpha J \left[ \epsilon + \sqrt{q_s} z \right]} \quad (192)$$

To complete the derivation, we apply the Fisher information convolution inequality (171) (see Appendix B.2) which implies that

$$J \left[ s^0 + \sqrt{q_d} z \right] \leq \frac{J \left[ s^0 \right]}{1 + q_d J \left[ s^0 \right]}, \quad (193)$$

substituting this into (186) yields
\[ q_s \geq q_d - q_d^2 J \left[ s^0 + \sqrt{q_d z} \right] \geq \frac{q_d}{1 + q_d J \left[ s^0 \right]}, \tag{194} \]

and since the RHS is monotonically increasing in \( q_d \), it is lower bounded by the lowest allowed value of \( q_d \) which is bounded by (190) so that

\[ q_s \geq \frac{1}{\alpha J \left[ \epsilon + \sqrt{q_s z} \right] + J \left[ s^0 \right]}, \tag{195} \]

**Result 13. Noise-free inference**

One application of the high dimensional regularized bound above is that it produces a lower bound on the required density of samples \( \alpha \) to achieve some permissible mean squared error per parameter \( Q \) in the big data limit only in terms of the allowed error \( Q \) and the Fisher information of the noise \( J \left[ \epsilon \right] \) and parameters \( J \left[ s^0 \right] \).

\[ \alpha_r \geq \frac{1 + Q J \left[ \epsilon \right]}{J \left[ s^0 \right]} \left( \frac{1}{Q} - J \left[ s^0 \right] \right). \tag{196} \]

**Derivation.** —

Re-arranging (178) we find

\[ \alpha \geq \frac{1 - q_s J \left[ s^0 \right] - \sqrt{q_s z} q_s J \left[ s^0 \right]}{q_s J \left[ \epsilon \right]} \geq \frac{(1 - q_s J \left[ s^0 \right])(1 + q_s J \left[ \epsilon \right])}{q_s J \left[ \epsilon \right]}, \tag{197} \]

where the second inequality follows from (171). Differentiating the RHS of (197) with respect to \( q_s \), we find that it is a monotonically decreasing quantity since the Fisher information is positive. Thus, under the condition \( q_s \leq Q \) the strongest bound occurs at \( q_s = Q \) and is

\[ \alpha_r \geq \max_{q_s \leq Q} \left[ \frac{(1 - q_s J \left[ s^0 \right])(1 + q_s J \left[ \epsilon \right])}{q_s J \left[ \epsilon \right]} \right] = \left[ \frac{(1 - Q J \left[ s^0 \right])(1 + Q J \left[ \epsilon \right])}{Q J \left[ \epsilon \right]} \right]. \tag{198} \]

Note that this inequality reveals how both the Fisher information of the noise and signal contributed to a lower bound on the requisite measurement density to achieve a prescribed error.

### 4.3 Lower bound on noise-free inference

**Result 14.** In the noise free regime (\( \epsilon \to 0 \)), we find the following lower bound on the MSE \( q_s \)

\[ q_s \geq q_s^{\text{MMSE}}(q_d), \quad \text{where} \quad q_d = \frac{q_s}{\alpha}, \tag{199} \]

and in terms of Fisher information,

\[ q_s \geq \frac{\alpha(1 - \alpha)}{J \left[ s^0 + \sqrt{\frac{Q}{\alpha} z} \right]} \geq \frac{1 - \alpha}{J \left[ s^0 \right]}. \tag{200} \]

**Derivation.** —

Note that the Result can be derived directly from the regularized bound (176), which simplifies in the noise-free limit \( \epsilon \to 0 \) so that \( q_d = \frac{q_s}{\alpha} \) and

\[ q_s \geq q_s^{\text{MMSE}}(q_d). \tag{201} \]

From the relation between scalar Fisher information and MMSE (see Appendix B.2), the above equation may be written as

\[ q_s \geq q_d(1 - q_d J \left[ s^0_{q_d} \right]) = \frac{q_s}{\alpha}(1 - \frac{q_s}{\alpha} J \left[ s^0_{q_d} \right]). \tag{202} \]

By dividing both sides of this equation by \( q_s \) and rearranging, we arrive at the inequality

\[ q_s \geq \frac{\alpha(1 - \alpha)}{J \left[ s^0_{q_d} \right]} = \frac{\alpha(1 - \alpha)}{J \left[ s^0 + \sqrt{\frac{Q}{\alpha} z} \right]}. \tag{203} \]

From the convolutional Fisher inequality (see Appendix B.2) we know
\[
\frac{1}{J[s^0_{\alpha d}]} \geq q_d + \frac{1}{J[s^0]},
\]

substituting this into inequality into the previous equation, after some rearrangement, yields

\[q_s \geq \frac{1 - \alpha}{J[s^0]}.
\]

4.4 Example: bounds on compressed sensing

**Result 15.** For noise free compressed sensing, where the signal \(s^0\) is drawn from a distribution which is \(f\)-sparse (only a fraction \(f\) of its components are non-zero), we find that the optimal convex M-estimator requires a number of measurements \(N\) satisfying

\[N \geq fP.
\]

to achieve perfect reconstruction to be possible in the asymptotic limit.

Consider the compressed sensing problem: the parameters being estimated are \(f\)-sparse so that a fraction \(f\) of them are non-zero. In this scenario, the bound from the previous result becomes

\[
\lim_{q_s \to 0} J[s^0 + \sqrt{\frac{\alpha}{q_s}}z] = \alpha \frac{1 - f}{q_s} + O(1).
\]

Substituting this into (203) we find that in the limit of perfect reconstruction, we have

\[\alpha \geq f.
\]

In other words we must have,

\[
\frac{N}{P} \geq f,
\]

so the number of measurements required is greater than the size of signal multiplied by the fraction of non-zero elements for an M-estimator to achieve perfect reconstruction in the asymptotic limit.
5 Optimal inference

5.1 Unregularized inference

Result 16. In the case of unregularized regression ($\sigma = 0$) and log-concave noise probability density $P_\epsilon$ (or equivalently convex noise energy $E_\epsilon = -\log P_\epsilon$), we find that the loss function with minimal MSE is

$$\rho^{opt} = -M_{q_s}[\log P_{\epsilon q_s}].$$

(210)

$M_\lambda[f]$ is a Moreau envelope defined as

$$M_\lambda[f](x) = \min_y \left[ \frac{(x-y)^2}{2\lambda} + f(y) \right].$$

(211)

(see Appendix C.1) and $q_s$ is the minimal solution satisfying

$$q_s = \frac{1}{\alpha J[\epsilon + \sqrt{q_s}z]}.$$

(212)

The distribution of signal and noise estimates are obtained through

$$\hat{s} = \sqrt{q_s}z + s^0,$$

(213)  $$\hat{\epsilon} = \epsilon^{MMSE}(\epsilon_{q_s}) = \int \epsilon P(\epsilon|\epsilon_{q_s}) d\epsilon,$$

(214) with probabilities $P(z)$ and $P(\epsilon_{q_s})$ respectively where $z$ is is a unit gaussian random variable. The MSE per component in estimating the noise is

$$q_\epsilon = q_\epsilon^{MMSE}(q_s) = \left\langle \left( (\epsilon + \sqrt{q_s}z) - \epsilon \right)^2 \right\rangle_{\epsilon, z}.$$

(215)

The training and test errors are

$$E^{\text{train}} = \sigma_\epsilon^2 - q_\epsilon = \sigma_\epsilon^2 - q_\epsilon^{MMSE}(q_s), \quad E^{\text{test}} = \sigma_\epsilon^2 + q_s.$$  

(216)

**Derivation of the Algorithm**

For ease of notation we will write $f = P_\epsilon$ and $\epsilon_q = \epsilon + \sqrt{q_s}$, where $z$ is a unit gaussian random variable, so that $\epsilon_q$ is a random variable drawn from pdf $f_q$. Our goal is to minimize the error $q_\epsilon$ by optimizing $\rho$ under the unregularized constraints from Result 4 using the method of lagrange multipliers and calculus of variations. Under the definition $r(x) = \lambda_\rho M_\lambda(\rho)'(x)$, the unregularized constraints (68) and (69) are

$$\alpha \int f_{\epsilon_q}(y)r(y)dy - q_s = 0,$$

(217)

$$\int f_{\epsilon_q}(y)r'(y)dy - \frac{1}{\alpha} = 0.$$  

(218)

Introducing multipliers $\lambda$ and $\mu$, our lagrangian may be written

$$L = q_s + \lambda \left[ \int f_{\epsilon_q}(y)r'(y)dy - \frac{1}{\alpha} \right] + \mu \left[ \int f_{\epsilon_q}(y) (r(y))^2 dy - \frac{q_s}{\alpha} \right].$$

(219)

The strategy we employ is to first find the optimal $r$, and second to derive $\rho^{opt}$ from $r$. Calculus of variations may be applied if we combine the integrals in the Lagrangian so that

$$L = q_s - \lambda \frac{1}{\alpha} - \mu \frac{1}{\alpha} q_s + \int S(r(y), r'(y))dy,$$

(220)

where

$$S(r, r') = (\lambda r' + \mu r^2) f_{\epsilon_q}.$$  

(221)

Defining $\delta L$ to be the variational derivative of $L$ (induced by varying $r$ by a small change $\delta r$). Taylor expanding the action $S$ in the integrand gives:

$$\delta L = \int \left( \frac{\partial S}{\partial r} \delta r + \frac{\partial S}{\partial r'} \delta r' \right) dy.$$  

(222)

Integrating the second term by parts gives
\[\delta L = \int \left( \frac{\partial S}{\partial r} - \frac{d}{dy} \frac{\partial S}{\partial r'} \right) \delta r dy. \quad (223)\]

For stationarity to hold for all \( \delta r \) we thus require the Euler-Lagrange (E-L) equation to hold, i.e.

\[\frac{\partial S}{\partial r} = \frac{\partial}{\partial y} \frac{\partial S}{\partial r'}. \quad (224)\]

Inserting (221) into (224) yields

\[2\mu r f_{\rho} = \lambda f_{\rho}', \quad (225)\]

Re-arranging,

\[r(y) = \frac{d}{dy} \frac{\lambda}{2\mu} \log(f_{\rho}(y)). \quad (226)\]

To solve for \( \frac{\lambda}{2\mu} \), we extremize the Lagrangian with respect to other variables after substituting the form of (225) into (219) to find

\[L = q_s - \frac{\lambda}{\alpha} - \frac{1}{\alpha} \mu q_s - \frac{\lambda^2}{4\mu} \int \left( f_{\rho}' \right)^2 f_{\rho}. \quad (227)\]

We can now use the fact that \( J[\epsilon + \sqrt{q_s}z] = \int \left( f_{\rho}' \right)^2 f_{\rho} \) to simplify notation. Now, extremizing \( L \) with respect to \( \lambda \) gives \( \frac{\lambda}{2\mu} = -\frac{1}{\alpha} J[\epsilon + \sqrt{q_s}z] \), and with respect to \( \mu \) gives (221). Combining these results we find \( \frac{\lambda}{2\mu} = -q_s \). Note that the zero solution can be neglected since \( \frac{\lambda}{2\mu} = -\frac{1}{\alpha} J[\epsilon + \sqrt{q_s}z] \). Substituting into \( L \) gives

\[L = q_s + \frac{\mu}{\alpha J[\epsilon + \sqrt{q_s}z]} \left( \frac{1}{\alpha} - q_s J[\epsilon + \sqrt{q_s}z] \right). \quad (228)\]

Because \( \mu \) is still a free parameter, we choose \( q_s \) to be the minimal value satisfying

\[J[\epsilon + \sqrt{q_s}z] q_s = \frac{1}{\alpha}. \quad (229)\]

Now we can substitute \( \frac{\lambda}{2\mu} = -q_s \) into (226) yields

\[r(x) = -q_s \frac{d}{dx} \log(f_{\rho}(x)). \quad (230)\]

Thus, from the definition of \( r(x) \),

\[\frac{d}{dx} \lambda_{\nu} M_{\lambda_{\nu}}[\rho^{\text{opt}}](x) = -q_s \frac{d}{dx} \log(f_{\rho}(x)). \quad (231)\]

Integrating both sides of this equation, we find that

\[M_{\lambda_{\nu}}[\rho^{\text{opt}}](x) = -q_s \frac{d}{dx} \log(f_{\rho}(x)) + b, \quad (232)\]

where \( b \) is an arbitrary constant. We now invert the Moreau envelope, using the property that for convex \( f \), \( M_{\lambda_{\nu}}[f] = g \) implies \( f = -M_{\lambda_{\nu}}[-g] \) (derived in Appendix C.2) yields

\[\rho^{\text{opt}} = -M_{\lambda_{\nu}}[\frac{q_s}{\lambda_{\nu}} \log(f_{\rho}) - b] = -M_{\lambda_{\nu}}[\frac{q_s}{\lambda_{\nu}} \log(f_{\rho})] + b. \quad (233)\]

Because additive constants will not effect on the M-estimation algorithm, we can write this as

\[\rho^{\text{opt}}(x) = -M_{\lambda_{\nu}}[\frac{q_s}{\lambda_{\nu}} \log(f_{\rho})](x) = -\min_y \left[ \frac{q_s}{\lambda_{\nu}} \log(f_{\rho}(y)) + \frac{(x-y)^2}{2\lambda_{\nu}} \right] = -\frac{q_s}{\lambda_{\nu}} M_{\lambda_{\nu}}[\log(f_{\rho})](x). \quad (234)\]

The unregularized RS equations from Result 4 have an additional symmetry that we can exploit, the equations are unchanged under the transform \( \rho \rightarrow c \rho \) and \( \lambda_{\nu} \rightarrow \frac{1}{c} \lambda_{\nu} \), where \( c \) is a positive real number. Since we can re-scale the loss function without changing the optimal of the inference algorithm, we simplify the optimal loss function by choosing \( \lambda_{\nu} = q_s \) so that

\[\rho^{\text{opt}} = -M_{q_s}[\log(f_{\rho})]. \quad (235)\]

Log-concavity of the noise distribution ensures that \( \rho^{\text{opt}} \) is convex (see C.3 for more details) so that the RS assumptions hold.
Result 17. Given the loss function form:

\[ \rho = -\mathcal{M}_q[\log P_{\epsilon_q}]. \]  

The proximal map of \( \rho \) becomes:

\[ \mathcal{P}_q[\rho](\epsilon_q) = \epsilon_{\text{MMSE}}(\epsilon_q). \]  

We first invert the Moreau envelope in (236) as explained in Appendix C.2, which yields

\[ \mathcal{M}_q[\rho](x) = -\log P_{\epsilon_q}. \]  

We want to compute the proximal map of \( \rho \), which is related to the Moreau envelope by the identity (401)

\[ \mathcal{P}_q[\rho](x) = x - q\mathcal{M}_q[\rho]'(x). \]

To calculate the second term, we differentiate the form of the Moreau envelope and multiply by a constant:

\[ -q\mathcal{M}_q[\rho]'(x) = q \frac{d}{dx} \log P_{\epsilon_q}(x) = q \frac{d}{dx} \int \frac{1}{\sqrt{2\pi q}} e^{-\frac{(x - y)^2}{2q}} P_q(y) dy = \int \frac{(y - x)}{\sqrt{2\pi q}} e^{-\frac{(y - x)^2}{2q}} P_q(y) dy = \int (\epsilon - \epsilon_{q_s}) P(\epsilon|\epsilon_{q_s}) d\epsilon. \]

The final equality follows from Bayes rule: we can derive an expression for the prior on noise \( \epsilon \) given a version corrupted by gaussian noise \( \epsilon_q \):

\[ P(\epsilon|\epsilon_q) = \frac{P(\epsilon_q|\epsilon)P(\epsilon)}{P(\epsilon_q)} = \frac{1}{\sqrt{2\pi q}} e^{-\frac{(\epsilon - \epsilon_q)^2}{2q}} P(\epsilon). \]

It then follows that

\[ \mathcal{P}_q[\rho](\epsilon_{q_s}) = \epsilon_{q_s} - q\mathcal{M}_q[\rho]'(\epsilon_{q_s}) = \int \epsilon P(\epsilon|\epsilon_{q_s}) d\epsilon = \epsilon_{\text{MMSE}}(\epsilon_{q_s}), \]

which yields the desired result that the proximal map becomes an MMSE estimate.

Distribution of estimates
We let \( q_s = q_s^{\text{opt}} \) be the optimal MSE for the remainder of this section for notational simplicity. We can use Result 17 to compute the distribution (67) of \( \epsilon \) in the unregularized optimal case where we can choose \( \lambda_\rho = q_s \) so that,

\[ \epsilon_{\text{opt}}(\epsilon_{q_s}) = \mathcal{P}_{q_s}[\rho^{\text{opt}}](\epsilon_{q_s}) = \mathcal{P}_{q_s}[\rho^{\text{opt}}](\epsilon_{q_s}) = \epsilon_{\text{MMSE}}(\epsilon_{q_s}). \]

MSE of noise estimation
We define \( q_s \) to be the per component MSE of an algorithm in estimating noise: \( q_s = \langle (\hat{\epsilon} - \epsilon)^2 \rangle \). As we showed in (83), in the optimal unregularized case,

\[ q_s = q_s \left( \frac{\alpha - 1}{\alpha} \right). \]

From (229), in the optimal case \( \alpha \) satisfies

\[ \alpha = \frac{1}{q_s J[\epsilon_{q_s}]]. \]

Combining these yields

\[ q_s = q_s - q_s^2 J[\epsilon_{q_s}] = q_s^{\text{MMSE}}(q_s). \]

The final equality follows from the relation between scalar Fisher information and MMSE given in Appendix B.4.

Training and test errors
Next we compute the training and test error. The test error is simply

\[ \mathcal{E}_{\text{test}}^{\text{opt}} = q_s + \sigma_{\epsilon}^2 = \frac{1}{\alpha J[\epsilon_{q_s}]} + \sigma_{\epsilon}^2. \]

We can also compute the training error
If we first compute the $q_e^{\text{MMSE}}(q_s)$,

$$q_{e}^{\text{MMSE}}(q_s) = \langle \langle \int (\varepsilon - \hat{e}^{\text{opt}}(q_s))^2 P(\varepsilon|q_s) d\varepsilon \rangle \rangle_{q_s} = \langle \langle \int e^2 P(\varepsilon|q_s) \rangle \rangle_{q_s} - \langle \langle (\hat{e}^{\text{opt}}(q_s))^2 \rangle \rangle_{q_s} = \sigma_e^2 - \mathcal{E}_{\text{opt}}^2, \quad (249)$$

thus

$$\mathcal{E}_{\text{opt}}^2 = \sigma_e^2 - q_e^{\text{MMSE}}(q_s). \quad (250)$$

### 5.2 Regularized Inference

**Result 18.** Under the assumption of log-concave probability densities $P_e$ and $P_s$, we find the optimal loss and regularizer pair to be:

$$\rho^{\text{opt}} = -M_{q_s}[\log (P_{q_s})], \quad \sigma^{\text{opt}} = -M_{q_d}[\log (P_{q_d})], \quad (251)$$

and $q_s$, $q_d$ are chosen such that $q_s$ is the minimum non-negative order parameter satisfying

$$q_d = \frac{1}{\alpha J[\varepsilon + \sqrt{q_s}]}, \quad q_s = q_d (1 - q_d J[s^0 + \sqrt{q_d}]) = q_s^{\text{MMSE}}(q_d). \quad (252)$$

The measurement deviations and parameter estimates take on a value

$$\hat{e} = \hat{e}^{\text{MMSE}}(q_s) = \int eP(e|q_s) d\varepsilon, \quad (253)$$
$$\hat{s} = \hat{s}^{\text{MMSE}}(q^0_{q_d}) = \int sP(s|q^0_d) ds, \quad (254)$$
with probability $P(s^0_{q_d})$, $P(e|q_s)$, moreover the MSE of the optimal $M$-estimator in estimating $\varepsilon$ is

$$q_e = \langle \langle (\hat{e}(q_s) - \varepsilon)^2 \rangle \rangle = q_e^{\text{MMSE}}(q_s). \quad (255)$$

The corresponding test and training errors are:

$$\mathcal{E}_{\text{test}}^2 = \sigma_e^2 + q_s = \sigma_e^2 + q_s^{\text{MMSE}}(q_d), \quad \mathcal{E}_{\text{train}}^2 = \sigma_e^2 - q_e = \sigma_e^2 - q_e^{\text{MMSE}}(q_s), \quad (256)$$

#### Derivation of the Algorithm

Our goal is to determine the optimal pair of functions $\rho, \sigma$, which we write as $\rho^{\text{opt}}, \sigma^{\text{opt}}$, to minimize the MSE of our estimation. We are optimizing simultaneously over regularizer and loss function to minimize $q_s$ under the RS constraints (70 - 73). To simplify notation, we will let $g = P_s$ and noise function $f = P_e$, and let $f_q$ and $g_q$ be the probability densities of random variables $e + \sqrt{q_s} z$ and $s^0 + \sqrt{q_d} z$, respectively for independent unit gaussian random variables $z_e, z_s$. Under the definition $r_{\rho}(x) = \lambda_{\rho} M_{\lambda_{\rho}}[\rho]'(x)$ and $r_{\sigma}(x) = \lambda_{\sigma} M_{\lambda_{\sigma}}[\sigma]'(x)$, we write these constraints as

$$\frac{q_d}{\lambda_{\rho}} - \frac{\alpha}{\lambda_{\rho}} \int f_q(y)r_{\rho}(y)^2 dy = 0, \quad (257)$$
$$q_s - q_d = \int g_{q_s}(y) \left( r_{\sigma}(y)^2 - 2 q_d r'_{\sigma}(y) \right) dy = 0 \quad (259)$$
$$\frac{1}{\lambda_{\sigma}} - \frac{\alpha}{\lambda_{\rho}} \int f_q(y)r_{\rho}'(y) dy = 0, \quad (258)$$
$$\lambda_{\rho} - \lambda_{\sigma} + \lambda_{\sigma} \int g_{q_s}(y)r_{\sigma}'(y) dy = 0 \quad (260)$$

Thus, we may define a Lagrangian for minimizing $q_s$ under the constraints (257,258,259,260) as

$$L = q_s + \gamma_1 \left[ \frac{q_d}{\lambda_{\rho}} - \frac{\alpha}{\lambda_{\rho}} \int f_q(y)r_{\rho}(y)^2 dy \right] + \gamma_2 \left[ \frac{1}{\lambda_{\sigma}} - \frac{\alpha}{\lambda_{\rho}} \int f_q(y)r_{\rho}'(y) dy \right]$$
$$\gamma_3 \left[ q_s - q_d - \int g_{q_s}(y) \left( r_{\sigma}(y)^2 - 2 q_d r'_{\sigma}(y) \right) dy \right] + \gamma_4 \left[ \lambda_{\rho} - \lambda_{\sigma} + \lambda_{\sigma} \int g_{q_s}(y)r_{\sigma}'(y) dy \right] \quad (261)$$

where we have introduced lagrange parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$. To apply calculus of variation, we re-write the lagrangian as

$$L = q_s + \frac{\gamma_1 q_d}{\lambda_{\rho}^2} + \frac{\gamma_2}{\lambda_{\sigma}} + \gamma_3 q_s - \gamma_3 q_d + \gamma_4 \lambda_{\rho} - \gamma_4 \lambda_{\sigma} + \int (S_{\rho}(y) + S_{\sigma}(y)) dy, \quad (262)$$
defining
\[ S_\rho(y) = -\alpha \frac{f_q(y)}{\lambda_\rho} \left( \gamma_1 r_\rho(y)^2 + \gamma_2 r'_\rho(y) \right) \] (263)
\[ S_\sigma(y) = -g_{q_d}(y) \left( \gamma_3 r_\sigma(y)^2 - (\gamma_4 \lambda_\sigma + 2 \gamma_3 q_d) r'_\sigma(y) \right). \] (264)

Requiring stationarity in both \( \delta r_\rho \) and \( \delta r_\sigma \) leads to the pair of E-L equations:
\[
\frac{\partial S_\rho(r_\rho, r'_\rho; y)}{\partial r_\rho} = \frac{d}{dy} \frac{\partial S_\rho(r_\rho, r'_\rho; y)}{\partial r'_\rho}, \tag{265}
\]
\[
\frac{\partial S_\sigma(r_\sigma, r'_\sigma; y)}{\partial r_\sigma} = \frac{d}{dy} \frac{\partial S_\sigma(r_\sigma, r'_\sigma; y)}{\partial r'_\sigma}, \tag{266}
\]
which lead to expressions \( r_\rho \) and \( r_\sigma \):
\[
r_\rho(y) = \lambda_\rho \frac{\gamma_2}{2 \gamma_1} \frac{d}{dy} \log(f_q(y)) \tag{267}
\]
\[
r_\sigma(y) = - \left( \frac{\gamma_4 \lambda_\sigma}{2 \gamma_3} + q_d \right) \frac{d}{dy} \log(g_{q_d}(y)) \tag{268}
\]
Substituting this form of \( r_\rho \) and \( r_\sigma \) into (263) and (264), and expressing the results in terms of the Fisher information, we find
\[
\int S_\rho(y) dy = \frac{\alpha \gamma_2^2}{4 \gamma_1} J \left[ \epsilon + \sqrt{q_d} z \right] \tag{269}
\]
\[
\int S_\sigma(y) dy = \gamma_3 \left( \frac{\gamma_4 \lambda_\sigma}{2 \gamma_3} + q_d \right)^2 J \left[ s^0 + \sqrt{q_d} z \right]. \tag{270}
\]
Thus, the Lagrangian (262) becomes
\[
L = \alpha \lambda_\rho q_d + \frac{\gamma_2}{\lambda_\sigma} + \gamma_3 q_d - \gamma_3 q_s - \gamma_4 \lambda_\rho - \gamma_4 \lambda_\sigma + \alpha \frac{\gamma_2^2}{4 \gamma_1} J \left[ \epsilon + \sqrt{q_d} z \right] + \gamma_3 \left( \frac{\gamma_4 \lambda_\sigma}{2 \gamma_3} + q_d \right)^2 J \left[ s^0 + \sqrt{q_d} z \right]. \tag{271}
\]
The optimization problem may be further simplified by extremizing with respect to \( \gamma_1 \) and \( \gamma_2 \), which yields
\[
\frac{q_d}{\lambda_\rho} = \frac{\alpha \gamma_2^2}{4 \gamma_1} J \left[ \epsilon + \sqrt{q_d} z \right] \tag{272}
\]
\[
\frac{q_d}{\lambda_\sigma} = - \frac{\gamma_2}{2 \gamma_1} J \left[ \epsilon + \sqrt{q_d} z \right] \tag{273}
\]
respectively. By combining these, we arrive at the constraints:
\[
\frac{\gamma_2}{2 \gamma_1} = - \frac{q_d}{\lambda_\rho} \tag{274}
\]
\[
q_d = \frac{1}{\alpha J} \left[ \epsilon + \sqrt{q_d} z \right] \tag{275}
\]
If we next extremize with respect to \( \lambda_\rho \), we find that \( \gamma_4 = 0 \). This suggests that constraint related to \( \gamma_4 \) does not impact the minimization problem. If we extremize with respect to \( \gamma_3 \) and \( \gamma_4 \) we find:
\[
q_s - q_d + q_\rho^2 J \left[ s^0 + \sqrt{q_d} z \right] = 0 \tag{276}
\]
\[
\lambda_\rho - \lambda_\sigma = q_d \lambda_\sigma J \left[ s^0 + \sqrt{q_d} z \right] \tag{277}
\]
Combining (276) and (277) yields an equation determining one order parameter given the other 3:
\[
q_s = q_d - q_\rho^2 J \left[ s^0 + \sqrt{q_d} z \right]. \tag{278}
\]
It follows that the optimization problem may be written in terms of only \( q_d, q_s \) as a minimum of \( q_s \) subject to constraints (275),(276). By defining
\[
q_d(q_s) = \frac{1}{\alpha J} \left[ \epsilon + \sqrt{q_d} z \right], \tag{279}
\]
the minimum MSE \( q_s^{opt} \) is the minimum value of \( q_s \) satisfying (276):
\[
q_s = q_d - q_\rho^2 J \left[ s^0 + \sqrt{q_d} z \right]. \tag{280}
\]
For convenience, we define \( q_s^{opt} = q_d(q_s^{opt}) \). Note that the form of (280) implies that \( q_s \) is the MMSE of reconstructing \( s^0 \) from \( s^0 + \sqrt{q_d} z \) where \( z \) is unit gaussian noise (see Appendix B.4 for a relationship between Fisher Information and MMSE).

To find the form of the optimal \( \rho, \sigma \), we first combine (274) and (278) to reduce (267) and similarly (268) to show
\[
\frac{d}{dy} M_{\lambda_\rho} [ \rho^{opt} ](y) = - \frac{q_\rho^{opt}}{\sigma_\rho} \frac{d}{dy} \log(f_q(y^{opt})), \tag{281}
\]
\[
\frac{d}{dy} M_{\lambda_\sigma} [ \sigma^{opt} ](y) = - \frac{q_\rho^{opt}}{\lambda_\sigma} \frac{d}{dy} \log(g_{q_d^{opt}}(y)), \tag{282}
\]
We can use the property derived in Appendix C.2 that \( M_{\lambda} [ f ] = g \) implies \( f = -M_{\lambda} [ -g ] \) under the assumption that \( f \) and \( g \) are convex along with the fact that the constants introduced by integrating both sides of the above equation will not effect the M-estimator, to write
\[
\rho^{opt} = -M_{\lambda_\rho} \left[ \frac{q_\rho^{opt}}{\lambda_\rho} \log(f_q^{opt}(y)) \right], \tag{283}
\]
\[
\sigma^{opt} = -M_{\lambda_\sigma} \left[ \frac{q_\rho^{opt}}{\lambda_\sigma} \log(g_{q_d^{opt}}) \right]. \tag{284}
\]
Or equivalently,
\[ \rho^{\text{opt}} = -\frac{q_s^{\text{opt}}}{\lambda_p} \mathcal{M}_{q_s^{\text{opt}}} \left[ \log(f_{q_s^{\text{opt}}}) \right] \quad (285) \]
\[ \sigma^{\text{opt}} = -\frac{q_d^{\text{opt}}}{\lambda_\sigma} \mathcal{M}_{q_d^{\text{opt}}} \left[ \log(g_{q_d^{\text{opt}}}) \right]. \quad (286) \]

The regularized RS equations from Result 3 have an additional symmetry that we can exploit, the equations are unchanged under the transform \((\rho, \sigma) \rightarrow (c\rho, c\sigma)\) with \((\lambda_p, \lambda_\sigma) \rightarrow \left(\frac{1}{c}\lambda_p, \frac{1}{c}\lambda_\sigma\right)\) where \(c\) is a positive real number. Since we can multiply \(\sigma^{\text{opt}}\) and \(\rho^{\text{opt}}\) by the same positive constant without effecting the results of the inference algorithm, we simplify the optimal M-estimator by rescaling such that \(\lambda_p = q_s^{\text{opt}}\), which from (278) implies \(\lambda_\sigma = q_d^{\text{opt}}\), so that
\[ \rho^{\text{opt}} = -\mathcal{M}_{q_s^{\text{opt}}} \left[ \log(f_{q_s^{\text{opt}}}) \right], \quad (287) \]
\[ \sigma^{\text{opt}} = -\mathcal{M}_{q_d^{\text{opt}}} \left[ \log(g_{q_d^{\text{opt}}}) \right]. \quad (288) \]

Log-concavity of the noise and parameter distributions ensures that \(\rho^{\text{opt}}\) and \(\sigma^{\text{opt}}\) are both convex (see C.3 for more details).

**Distribution of Estimates**

For ease of notation let \(q_s = q_s^{\text{opt}}\) and \(q_d = q_d^{\text{opt}}\) for the remainder of this section. As we derived previously (63,64), the distribution measurement deviation \(\hat{\epsilon} = y - X\hat{s}\) and of estimates \(\hat{s}\) are
\[ \hat{e}^{\text{opt}}(\epsilon_{q_s}) = \mathcal{P}_{\lambda_p}[\rho^{\text{opt}}](\epsilon_{q_s}), \quad (289) \]
\[ \hat{s}^{\text{opt}}(s_{q_d}^0) = \mathcal{P}_{\lambda_\sigma}[\sigma^{\text{opt}}](s_{q_d}^0). \quad (290) \]

Since we can set \(\lambda_p = q_s, \lambda_\sigma = q_d\) in the optimal case, we can apply Result 17, and find that
\[ \hat{e}^{\text{opt}}(\epsilon_{q_s}) = \mathcal{P}_{q_s}[\rho^{\text{opt}}](\epsilon_{q_s}) = \hat{e}^{\text{MMSE}}(\epsilon_{q_s}), \quad (291) \]
\[ \hat{s}^{\text{opt}}(s_{q_d}^0) = \mathcal{P}_{q_d}[\sigma^{\text{opt}}](s_{q_d}^0) = \hat{s}^{\text{MMSE}}(s_{q_d}^0). \quad (292) \]

It also follows that
\[ q_s^{\text{MMSE}}(q_d) = \langle \langle (s^0 - \hat{s}^{\text{opt}})^2 P(s^0|s_{q_d}^0) ds^0 \rangle \rangle_{s_{q_d}^0} = \langle \langle (s^0 - \hat{s}^{\text{opt}})^2 \rangle \rangle_{s_{q_d}^0} = q_s, \quad (293) \]
so \(q_s\) is the MMSE of inferring \(s^0\) from \(s_{q_d}^0\).

**MSE of Noise Estimation**

As we showed in (82), in the optimal regularized case,
\[ q_e = q_s + \frac{q_s}{\alpha} \left[ \frac{q_d}{q_s} \left( \frac{\lambda_p}{\lambda_\sigma} - \frac{q_s}{q_d} \right)^2 - \frac{q_s}{q_d} \right]. \quad (294) \]

Choosing \(\lambda_p = q_s, \lambda_\sigma = q_d\) and substituting the form of \(q_d\) (279), this expression can be simplified as
\[ q_e = q_s - q_s^2 J[\epsilon_{q_s}] = q_e^{\text{MMSE}}(q_s). \quad (295) \]
See the relation between scalar Fisher information and MMSE in Appendix B.4.

**Training and Test Errors**

The test error is simply
\[ \mathcal{E}_\text{test} = \sigma_e^2 + q_s = \sigma_e^2 + q_s^{\text{MMSE}}(q_d). \quad (296) \]

From the definition of training error, we have that
\[ \mathcal{E}_\text{train} = \langle \langle \epsilon^2 \rangle \rangle_{\epsilon_{q_e}}, \quad \mathcal{E}_\text{opt}^{\text{train}} = \langle \langle (\hat{e}^{\text{opt}})^2 \rangle \rangle_{\epsilon_{q_e}}. \quad (297) \]

It then follows from (291) that
\[ q_e^{\text{MMSE}}(q_s) = \langle \langle (\epsilon - \hat{e}^{\text{opt}}(\epsilon_{q_s}))^2 P(\epsilon|\epsilon_{q_s}) d\epsilon \rangle \rangle_{\epsilon_{q_s}} = \langle \langle (\epsilon^2 P(\epsilon|\epsilon_{q_s})) \rangle \rangle_{\epsilon_{q_s}} - \langle \langle (\hat{e}^{\text{opt}}(\epsilon_{q_s}))^2 \rangle \rangle_{\epsilon_{q_s}} = \sigma_e^2 - \langle \langle (\hat{e}^{\text{opt}}(\epsilon_{q_s}))^2 \rangle \rangle_{\epsilon_{q_s}}. \quad (298) \]

Rearranging, and substituting the form of \(\mathcal{E}_\text{opt}^{\text{train}}\), it immediately follows that
\[ \mathcal{E}_\text{opt}^{\text{train}} = \sigma_e^2 - q_e^{\text{MMSE}}(q_s). \quad (299) \]
5.3 Analytic bounds on MSE

Result 19. The accuracy of unregularized $M$-estimation is invariant with respect to the true $s^0$ distribution and the optimal MSE obeys:

$$\frac{1}{\sigma^2 J[\epsilon]} \leq q_s^{\text{opt}} \leq \frac{1}{\alpha - 1}. \quad (300)$$

In the regularized case, the optimal MSE $q_s^{\text{opt}} = q_s^{\text{quad}}$ satisfies the bounds

$$\frac{1}{\sigma^2 J[s^0]} \bar{q}_s^{\text{quad}} \left( \alpha, \frac{J[s^0]}{J[\epsilon]} \right) \leq q_s^{\text{opt}} \leq \bar{q}_s^{\text{quad}} \left( \alpha, \frac{\sigma^2}{\sigma^2_s} \right), \quad (301)$$

where $\bar{q}_s^{\text{quad}}(\alpha, \Phi)$, the form of which is derived in (118), is the positive root of the quadratic

$$x^2 + (\Phi + \alpha - 1)x - \Phi = 0. \quad (302)$$

Note that under the condition of Gaussian signal and noise $J[s^0] = \frac{1}{\sigma^2}, J[\epsilon] = \frac{1}{\sigma^2},$ thus

$$q_s^{\text{opt}} = \bar{q}_s^{\text{quad}} \left( \alpha, \frac{\sigma^2}{\sigma^2_s} \right). \quad (303)$$

Derivation. — In the unregularized case, quadratic $M$-estimation (least-squares regression) has the form (116) which must upper bounds the optimal MSE. The high dimensional Cramer-Rao bound (163) lower bounds the MSE. The upper bound on the optimal regularized $M$-estimator follows from Result 9 for the optimal quadratic $M$-estimator, which must necessarily have an error greater than or equal to that of the optimal $M$-estimator. To derive our lower bound, we apply the convolutional Fisher inequality to the bounds derived for regularized $M$-estimation:

$$q_d \geq \frac{1}{\alpha J[\epsilon_q]}, \quad q_s \geq q_d \left( 1 - q_d J[s^0] \right) \geq q_d \left( 1 - \frac{q_d J[s^0]}{q_d J[s^0] + 1} \right). \quad (304)$$

Combining the previous two inequalities yields the desired lower bound.

5.4 Limiting cases

Result 20. Limiting Case for Optimal $M$-estimation: The form of the optimal $M$-estimator derived for log concave parameter and noise distributions further simplifies under various limits of the measurement density $\alpha$ and SNR:

<table>
<thead>
<tr>
<th>High SNR</th>
<th>SNR is $O(1)$</th>
<th>Low SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case $\alpha &gt; 1$</td>
<td>$q_s \approx \frac{1}{\alpha J[\epsilon_q]}$</td>
<td>$q_s \approx \frac{1}{\alpha J[\epsilon]}$</td>
</tr>
<tr>
<td>Case $\alpha \ll 1$</td>
<td>$q_s \approx \sigma^2 q_s^{\text{quad}} \left( \alpha, \frac{\sigma^2}{\sigma^2_s} \right)$</td>
<td>$q_s \approx \frac{1}{\sigma^2} + \alpha J[\epsilon_q]$</td>
</tr>
<tr>
<td>Case $\alpha \gg 1$</td>
<td>$q_s \approx \frac{1}{\alpha J[\epsilon]}$</td>
<td>$q_s \approx \frac{1}{\sigma^2} + \alpha J[\epsilon_q]$</td>
</tr>
</tbody>
</table>

Derivation. — We will choose $\sigma_s$ to be $O(1)$ and vary the scale of the noise $\sigma$ corresponding to different limits of SNR. We will also assume that $P_\epsilon, P_s$ have finite Fisher information when their variances are finite. From Result 18, in cases when the optimal $M$-estimator may be computed, MSE $q_s$ obeys the conditions:

$$q_d = \frac{1}{\alpha J[\epsilon_q]}, \quad q_s = q_d(1 - q_d J[s^0]). \quad (306)$$

25
High SNR ($\sigma_\epsilon \to 0$)

Case $\alpha > 1$: In the over-sampled regime, where $\alpha > 1$, even least-squares regression will achieve reconstruction with an error proportional to $\sigma_\epsilon$, as can be seen in (116). The optimal M-estimator performance has a strictly lower error than least-squares, thus $q_s \to 0$. It follows that $J[\epsilon_{q_s}]$ approaches infinity, so from (306) we know $q_d \to 0$. It follows from (307) that in this limit of small $q_s$ and $q_d$, $q_s \approx q_d$. Substituting this into (306) yields

$$q_s \approx \frac{1}{\alpha J[\epsilon_{q_s}]}.$$  \hfill (308)

Of course, we need $\sigma_\epsilon$ sufficiently small so that $q_s \approx 0$, and it is sufficient to bound $q_s$ from above with the error from least squares: $q_s^{LS} = \frac{\sigma_\epsilon^2}{\alpha - 1}$, we thus require lower noise when $\alpha$ is close to one such that $\sigma_\epsilon^2 \ll \alpha - 1$. A regularizer is not required to accurately estimate $q_s$ in this regime, so information about the parameter distribution is not useful and we find that the high dimensional unregularized bound is an accurate approximation of achievable MSE.

Case $\alpha \ll 1$: $q_s$ will be bounded above by $\sigma_\epsilon^2$ since it must outperform the trivial estimator which simply estimates zero for every parameter. $q_s$ will also be non-zero for finite $J[s_0]$, as can be seen from the lower bound from (301). It follows that for $\alpha \to 0$,

$$q_d = \frac{1}{\alpha J[\epsilon_{q_s}]} \to \infty.$$  \hfill (309)

In this limit, $s_0^{q_d}$ may be approximated as a gaussian with variance $q_d + \sigma_\epsilon^2$. Similarly $q_s$ is $O(1)$ and the variance of $\epsilon$ is small, so $\epsilon_{q_s}$ is also approximately gaussian and

$$J[\epsilon_{q_s}] \approx \frac{1}{q_s + \sigma_\epsilon^2}, \quad J[s_0^{q_d}] \approx \frac{1}{q_d + \sigma_\epsilon^2}.$$  \hfill (310)

Substituting into the optimal M-estimator conditions and solving for $q_s$ yields the result that the optimal M-estimator error is equivalent to that of the optimal quadratic loss-regularizer pair in this regime.

SNR is $O(1)$ ($\sigma_\epsilon = O(1)$)

Case $\alpha \gg 1$: In the the highly over-sampled regime, again even least squares regression approaches zero error, so that the optimal M-estimator has zero error $q_s \to 0$. Since $\alpha$ is large

$$q_d \approx \frac{1}{\alpha J[\epsilon]} \to 0.$$  \hfill (311)

It follows from (307) that for small $q_d$, $q_s \approx q_d$:

$$q_s \approx \frac{1}{\alpha J[\epsilon_{q_s}]} \approx \frac{1}{\alpha J[\epsilon]}.$$  \hfill (312)

where the final approximation follows from the fact that $q_s$ is small and $\sigma_\epsilon$ is order 1. Therefore this is a regime where the Cramer-Rao bound becomes tight and maximum likelihood becomes optimal.

Case $\alpha \ll 1$: It immediately follows from (306) that $q_d \to \infty$ which implies that $s_0^{q_d}$ can be approximated by a gaussian so (307) becomes

$$q_s \approx q_d \left(1 - \frac{q_d}{q_d + \sigma_\epsilon^2}\right) = \frac{1}{\alpha J[\epsilon_{q_s}]} + \frac{1}{\sigma_\epsilon^2}.$$  \hfill (313)

Low SNR ($\sigma_\epsilon \to \infty$)

Case $\alpha < \infty$: In this regime $q_d$ approaches infinity, which follows from our assumption that for large $\sigma_\epsilon$ the Fisher information of the noise approaches zero. For large $q_d$, we may approximate $s_0^{q_d}$ as a gaussian and recover (313). Since $q_s = O(1)$, it is very small compared to $\epsilon$, so that $\epsilon_{q_s} \approx \epsilon$:

$$q_s \approx \frac{1}{\alpha J[\epsilon] + \frac{1}{\sigma_\epsilon^2}}.$$  \hfill (314)

26
### 5.5 Optimal noise-free inference

**Result 21.** In the case of noise-free reconstruction (for $P_s$ log-concave) the optimal regularization function is

$$\sigma^{opt}(x) = -M_{q_s}[\log P_{q_d}](x),$$  \hfill (315)

and yields a MSE of $q_s$, where $q_s$ and $q_d$ satisfy

$$q_s = q_s^{MMSE}(q_d),$$  \hfill (316)  

$$q_d = \frac{q_s}{\alpha},$$ \hfill (317)

and the estimated parameters $\hat{s}$ of the M-estimator are distributed such that they take the value

$$\hat{s} = \hat{s}^{MMSE}(s_0^{q_d}),$$ \hfill (318)

with probability $P(s_0^{q_d})$.

This result follows from minimizing the MSE $q_s$ under the noise free error relations of Result 7. We have that $\lambda_\rho = \alpha \lambda_\sigma$, $q_s = \alpha q_d$, and by substituting this pair of identities into (259) and (260), we find

$$\alpha q_d - q_d - \int g_{q_d}(y) \left( r_\sigma (y)^2 - 2q_d r'_\sigma (y) \right) dy = 0, \hfill (319)$$

$$1 + \frac{1}{\alpha} \left( \int g_{q_d}(y) r'_\sigma (y) dy - 1 \right) = 0. \hfill (320)$$

If we then follow the same procedure of constructing a Lagrangian and performing calculus of variations, as in the previous section, we arrive at the form of $\sigma^{opt}$ above as well as the constraint on $q_s$:

$$q_s = \frac{\alpha}{J} \left[ 1 - \alpha \right] \left( s_0 + \sqrt{\frac{1}{\alpha} q_s^2} \right). \hfill (321)$$

In order to remove $\alpha$ from the above equation, we first re-arrange it to solve for $\alpha$:

$$\alpha = 1 - q_d J \left[ s_0^{q_d} \right]. \hfill (322)$$

Substituting the form of $\alpha$ into the preceding equation allows us to write $q_s$ as

$$q_s = q_d \left( 1 - q_d J \left[ s_0^{q_d} \right] \right) = q_s^{MMSE}(q_d), \hfill (323)$$

where the final equality follows from relation between Fisher information and MMSE in Appendix B.4. We know from the regularized RS equations that the parameter distribution is distributed as

$$\hat{s} = P_{\lambda_\sigma}[\sigma](s_0^{q_d}). \hfill (324)$$

Applying Result 17 to the form of $\sigma^{opt}$, and using the fact that we can set $\lambda_\sigma = q_d$ for optimal inference, the distribution of estimated parameters (318) immediately follows. Note that it also follows directly from the form of the estimated parameters that

$$q_s = \langle (\hat{s}^{MMSE}(s_0^{q_d}) - s_0^{q_d})^2 \rangle = q_s^{MMSE}(q_d). \hfill (325)$$

### 5.6 The limits of small and large smoothing of Moreau envelopes

**Result 22.** We derive the form of the optimal M-estimator in the limits of both small and large smoothing. This yields simplified forms for both optimal unregularized and regularized inference at high and low measurement densities. Letting $f_q$ denote the density of a pdf $f$ convolved with a zero mean gaussian of variance $q$, and defining $\mu_f, \sigma_f^2$ to be the mean and variance of $f$ respectively, we find:

$$\lim_{q \to 0} M_q[\log f_q](x) = \log f(x) \hfill (326)$$

$$\lim_{q \to \infty} M_q[\log f_q](x) = \frac{(x - \mu_f)^2}{2\sigma_f^2}. \hfill (327)$$

*A sufficient condition for (326) is that $\left| \frac{\partial f}{\partial x} \right|$ is bounded above. For details on the subgradient $\partial$ see e.g. [4]. A sufficient condition for (327) is finite moments/cumulants of $f$.  

27
\[
\mathcal{M}_q[\log f_q](x) = \min_y \left[ \log f_q(y) + \frac{(x-y)^2}{2q} \right] = \left[ \log f_q(y^*_q(x)) + \frac{(x-y^*_q(x))^2}{2q} \right]
\]

where \(y^*_q(x)\) is the minimizer of the Moreau envelop (the prox function), and thus satisfies

\[
y^*_q(x) = x + q \frac{\partial f_q(y^*_q(x))}{f_q(y^*_q(x))}.
\]

In the limit \(q \to 0\),

\[
\lim_{q \to 0} \mathcal{M}_q[\log f_q](x) = \lim_{q \to 0} \left[ \log f_q \left( x + q \frac{\partial f_q(y^*_q(x))}{f_q(y^*_q(x))} + q \frac{\left( \frac{\partial f_q(y^*_q(x))}{f_q(y^*_q(x))} \right)^2}{2} \right) \right] = \log f_0(x) = \log f(x),
\]

where the second equality follows from the assumption of bounded \(|\partial f|\) (so that \(\lim_{q \to 0} \frac{\partial f}{f} = 0\)). For the case where \(q \to \infty\),

\[
2\pi f_q(y) = \int dk e^{iky} e^{-\frac{q^2}{2} \log \tilde{f}(k)} = \int dk e^{iky} e^{-\frac{q^2}{2} + \log \hat{f}(k)} = \int dk e^{-q \left( \frac{k^2}{2} - i\frac{k\mu}{q} - \frac{i}{2} \log \hat{f}(k) \right)}.
\]

Thus, letting \(H(k) = \log \hat{f}(k)\),

\[
f_\infty(y) = \frac{1}{2\pi} \lim_{q \to \infty} \int dk e^{-q \left( \frac{k^2}{2} - i\frac{k\mu}{q} - \frac{i}{2} H(k) \right)}.
\]

Now, assuming \(f\) has finite cumulants, we can calculate the above function via a saddle point argument that

\[
f_\infty(y) = \lim_{q \to \infty} e^{-q \left( \frac{k^2}{2} - i\frac{k\mu}{q} - \frac{i}{2} H(k) \right)} \left[ \frac{1}{\sqrt{2\pi(q + \sigma^2)}} \right],
\]

where \(\hat{k}\) is the maximizer of the integrand. A perturbative expansion of \(\hat{k}\) gives

\[
\hat{k} = \frac{iy + H'(0)}{q} + \frac{H''(0)(iy + H'(0))}{q^2} + \ldots
\]

It is not hard to show that \(H(0) = 0\), \(H'(0) = -i\mu\), and \(H''(0) = -\sigma^2\), where \(\mu\) and \(\sigma\) are respectively the mean and variance of \(f\) and find

\[
f_\infty(y) = \lim_{q \to \infty} \frac{1}{\sqrt{2\pi(q + \sigma^2)}} e^{-\frac{(y - \mu)^2}{2(q + \sigma^2)}}, \quad \text{for } q \to \infty.
\]

Thus, the smoothed distribution tends to a gaussian, where we can approximate the original pdf \(f\) by a gaussian with identical variance and mean. We can now use this approximation of \(\zeta_q\) to compute \(\rho\) in the large \(q\) limit

\[
\rho_\infty(x) = \lim_{q \to \infty} \inf_y \left[ \frac{(y - \mu)^2}{2(q + \sigma^2)} + \frac{(x - y)^2}{2q} \right].
\]

After some manipulation we find

\[
\rho_\infty(x) = \frac{(x - \mu)^2}{2\sigma^2}.
\]

We use this result in the main paper to demonstrate that under suitable assumptions in the high dimensional regime the optimal unregularized loss approaches a quadratic, and the optimal regularized M-estimator has a regularization function approaching a square penalty.

These results can be applied to simplify optimal inference in various limits. For example in unregularized inference, as \(\alpha \to 1\), \(q_s \to \infty\), so the results of this section imply that the optimal loss function (210) simplifies to a quadratic in the under-sampled limit. Also, as \(\alpha \to \infty\), \(q_s \to 0\), implying that the optimal loss function simplifies to maximum-likelihood in the over-sampled limit.

In the case of regularized inference, as \(\alpha \to 0\), \(q_d \to \infty\) so that the optimal regularization function (251) approaches a quadratic. If we are additionally in the low noise (high SNR) limit, then \(q_s\) is \(O(1)\) while \(\epsilon\) is small so that smoothing dominates and the optimal loss also approaches a quadratic. In the over-sampled limit of \(\alpha \to \infty\), \(q_d \to 0\) and \(q_s = q_d - q_d^2 J \left[ s_{q_d}^0 \right] \approx q_d\), since the smoothing for both the loss and regular approach 0 at the same rate the optimal M-estimator (251) approaches MAP.
5.7 Comparison of optimal M-estimation with MMSE accuracy

In this section we derive replica theory predictions, using the finite temperature RS relations we derived in Result 1, to compute the Bayes-optimal MMSE. The MMSE estimator is simply the mean of the posterior distribution of the parameters given data \( y, X \) and one may compute this posterior distribution using Bayes rule:

\[
P(s^0|y, X) = \frac{P(y|s^0, X)P(s^0)P(X)}{P(y|X)P(X)} = \frac{P(y|s^0, X)P(s^0)}{P(y|X)}.
\]

(338)

It follows that under our assumption that noise and parameters are drawn iid from pdf \( P_e \) and \( P_s \) respectively, the posterior distribution takes the form:

\[
P(s^0|y, X) \propto \prod_{\mu=1}^{N} P_e(y_\mu - x_\mu \cdot s^0) \prod_{j=1}^{P} P_s(s_j^0).
\]

(339)

In the finite temperature regime, under the choice of unit temperature \( \beta = 1 \), \( \sigma = -\log P_s \), and \( \rho = -\log P_e \) we find that the energy distribution (4) will yield a Gibbs distribution precisely equivalent to the posterior of the parameters:

\[
P_G(s) = \frac{1}{Z(y, X)} e^{-\beta(\sum_{\mu=1}^{N} \rho(y_\mu - x_\mu \cdot s) + \sum_{j=1}^{P} \sigma(s_j))} \propto \prod_{\mu=1}^{N} P_e(y_\mu - x_\mu \cdot s) \prod_{j=1}^{P} P_s(s_j).
\]

(340)

It follows that averaging \( s \) over the Gibbs distribution \( \langle s \rangle \) yields the MMSE estimator and that \( \langle \langle (s) - s^0 \rangle \rangle \) is simply the MMSE. The finite temperature replica symmetric equations from Result 1 allows us to characterize the MMSE solution as a 1-dimensional MMSE inference problem. Under the definitions

\[
\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) = \frac{1}{P(\epsilon_{q_s})} \int \epsilon P_e(\epsilon) \frac{1}{\sqrt{2\pi q_s}} e^{-\frac{(\epsilon - \epsilon_{q_s})^2}{2q_s}} d\epsilon,
\]

(341)

\[
b_{\text{MMSE}}(s_{q_s}) = \frac{1}{P(s_{q_s})} \int s P_s(s) \frac{1}{\sqrt{2\pi q_d}} e^{-\frac{(s - s_{q_d})^2}{2q_d}} ds.
\]

(342)

In searching for solutions to the finite temperature RS relations in this setting, we make the ansatz that \( \Lambda_\rho = q_s, \Lambda_\sigma = q_d \), so that \( P_\rho \) and \( P_\sigma \) become posterior distributions of scalar Gaussian corrupted noise and corrupted signal \( P(\epsilon|x_{q_s}) \) and \( P(s|x_{q_d}) \) respectively. In this setting, the finite temperature RS relations thus reduce to:

\[
\frac{1}{q_d} = \frac{\alpha}{q_s} \langle \langle (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon_{q_s})^2 \rangle \rangle_{\epsilon_{q_s}} \quad (343) \quad q_s = q_{s_{\text{MMSE}}}(q_d)
\]

(345)

\[
\frac{1}{q_d} = \frac{\alpha}{q_s} (q_s - q_{s_{\text{MMSE}}}(q_s)) \quad (344) \quad q_s = q_{s_{\text{MMSE}}}(q_d).
\]

(346)

We now demonstrate that (343) is equivalent to (344). To this end,

\[
\langle \langle (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon_{q_s})^2 \rangle \rangle_{\epsilon_{q_s}} = \langle \langle (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon - \sqrt{q_s} z_s)^2 \rangle \rangle_{\epsilon, z_s} \quad (347)
\]

\[
= \langle \langle (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon)^2 \rangle \rangle - 2\sqrt{q_s} \langle \langle z_s (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon) \rangle \rangle_{\epsilon, z_s} + q_s \quad (348)
\]

From the definition of \( q_{s_{\text{MMSE}}}(q_s) \), we can rewrite the previous expression as

\[
\langle \langle (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon_{q_s})^2 \rangle \rangle_{\epsilon_{q_s}} = q_{s_{\text{MMSE}}}(q_s) + q_s - 2\sqrt{q_s} \langle \langle z_s (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon) \rangle \rangle_{\epsilon, z_s} \quad (349)
\]

To complete the derivation we simplify the final term in the previous expression, defining

\[
A = \sqrt{q_s} \langle \langle z_s (\hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) - \epsilon) \rangle \rangle_{\epsilon, z_s} = \sqrt{q_s} \langle \langle z_s \int \epsilon P(\epsilon|x_{q_s}) \frac{1}{\sqrt{2\pi q_s}} e^{-\frac{(\epsilon - \epsilon_{q_s})^2}{2q_s}} d\epsilon \rangle \rangle_{\epsilon, z_s}.
\]

(350)

Now, applying the identity \( \langle \langle z f(z) \rangle \rangle = \langle \langle f'(z) \rangle \rangle \) for a unit gaussian variable \( z \) yields

\[
A = \langle \langle \epsilon P(\epsilon|x_{q_s}) d\epsilon \rangle \rangle_{\epsilon, z_s} - \langle \langle (\epsilon) (\epsilon - \sqrt{q_s} z_s) \rangle \rangle_{\epsilon, z_s} = \langle \langle \epsilon^2 \rangle - \langle \epsilon \rangle^2 - \langle \epsilon \rangle \epsilon \rangle \rangle_{\epsilon, z_s} \quad (351)
\]

where \( \langle \epsilon \rangle = \int \epsilon P(\epsilon|x_{q_s}) d\epsilon = \hat{\epsilon}_{\text{MMSE}}(\epsilon_{q_s}) \). The expression for \( A \) can thus be simplified to

\[
A = \langle \langle \epsilon^2 \rangle - \langle \epsilon \rangle^2 \rangle \rangle_{\epsilon, z_s} = q_{\epsilon_{\text{MMSE}}}(q_s).
\]

(352)
Substituting this result into (343) yields (344). By applying the relationship between Fisher information and scalar MMSE (388), the finite temperature RS relations yield:

\[
q_{e\text{MMSE}}(q_s) = q_s - q_\sigma^2 J [\epsilon_q], \quad (353) \\
q_{s\text{MMSE}}(q_d) = q_d - q_\sigma^2 J [s_0^q]. \quad (354)
\]

Substituting these forms into (344) and (345) implies

\[
q_d = \frac{1}{\alpha J [\epsilon_q]}, \quad (355) \\
q_s = q_{s\text{MMSE}}(q_d) = q_d - q_\sigma^2 J [s_0^q]. \quad (356)
\]

This set of self-consistency equations for optimal Bayesian inference in the thermodynamic limit is consistent with previous MMSE predictions in the literature such as [5] (eqn. 12) which yields a set of self-consistency equations equivalent to ours under the choice of Gaussian noise, in which case (355) reduces to:

\[
q_d = \frac{\sigma^2 + q_s}{\alpha}.
\]

Remarkably, eqs. (355)-(356) exactly match the RS relations for optimal M-estimation (252), implying that the M-estimators we have derived achieve MMSE performance when the RS approximation is exact. A sufficient condition for the validity of RS is to assume log-concave noise and signal distributions.

A Review of Single Parameter Inference

A.1 Single parameter M-estimation

Consider the problem of inferring a single parameter \(s^0\) given measurements \(y_\mu\) for \(\mu = 1, ..., N\):

\[
y_\mu = s^0 + \epsilon_\mu, \quad (358)
\]

where the noise \(\epsilon_\mu\) is drawn iid from a probability density function \(P_\epsilon\). We apply an unregularized M-estimator to estimate \(s^0\):

\[
\hat{s} = \arg \min \sum_\mu \rho(y_\mu - s) = \arg \min \sum_\mu \rho(\epsilon_\mu - (s - s^0)). \quad (359)
\]

Thus \(\hat{s}\) is the solution to,

\[
\sum_\mu \rho'(\epsilon_\mu - (\hat{s} - s^0)) = 0. \quad (360)
\]

For an unbiased M-estimator, in the large \(N\) limit \(\hat{s} - s^0\) becomes small so that we can Taylor expand in this small quantity:

\[
\sum_\mu \rho'(\epsilon_\mu) - \sum_\mu \rho''(\epsilon_\mu)(\hat{s} - s^0) = 0. \quad (361)
\]

Rearranging and squaring both sides,

\[
\left(\sum_\mu \rho''(\epsilon_\mu)\right)^2 (\hat{s} - s^0)^2 = \left(\sum_\mu \rho'(\epsilon_\mu)\right)^2 - \sum_\mu \rho'(\epsilon_\mu)^2, \quad (362)
\]

Here we neglect the off diagonal terms in the final square, as the terms in the sum are uncorrelated random variables, so the average of the off diagonal terms will vanish. Thus, in the asymptotic large \(N\) limit, after averaging over the noise, we find

\[
q_s = (\hat{s} - s^0)^2 = \frac{\sum_\mu \rho''(\epsilon_\mu)^2}{\left(\sum_\mu \rho''(\epsilon_\mu)\right)^2} \to \frac{\langle \rho''(\epsilon) \rangle}{\langle \rho'(\epsilon) \rangle^2}, \quad (363)
\]

where we have define \(q_s\) to be the MSE of signal estimate.

A.2 Bounds on single parameter M-estimation

We can use this result on the form (363) of the MSE of an M-estimator to derive a bound on the MSE of M-estimation. We begin with the expression in the denominator and apply integration by parts,

\[
\left(\int \rho''(x)P_\epsilon(x)dx\right)^2 = \left(\int \rho'(x)P'_\epsilon(x)dx\right)^2 = \left(\int \left(\rho'(x)\sqrt{P_\epsilon(x)}\frac{P'_\epsilon(x)}{\sqrt{P_\epsilon(x)}}\right)dx\right)^2 \leq \int P_\epsilon(x)\rho'(x)^2dx \int \frac{P'_\epsilon(x)^2}{P_\epsilon(x)}dx, \quad (364)
\]
where the inequality follows from Cauchy-Schwartz. The final integral is simply the Fisher information from (382). This bound will turn out to be a special case of a more general bound known as the Cramer-Rao bound, that applies any unbiased inference problem. Thus, our analytic asymptotic expression (363) for MSE is bounded as

\[ q_s = \frac{\langle \langle \rho'(\epsilon)^2 \rangle \rangle}{\langle \langle \rho''(\epsilon) \rangle \rangle^2} \geq \frac{1}{\int \frac{P'_s(x)^2}{P_s(x)} \, dx} = \frac{1}{J[\epsilon]}. \]  

(365)

The Cauchy-Schwartz inequality is saturated when the functions in the inner product (364) are proportional so that

\[ \frac{\rho'(x)}{\sqrt{P_s(x)}} \approx \frac{P'_s(x)}{\sqrt{P_s(x)}}. \]  

(366)

It follows that \( \rho(x) = -\log P_s(x) \) is the asymptotically optimal M-estimator and achieves the optimal minimum MSE amongst all unbiased estimators.

### A.3 Asymptotic analysis of single parameter Bayesian MMSE estimation

Suppose we are given \( N \) noisy measurements indexed by \( \mu = 1, ..., N \) of an unknown signal \( s^0 \):

\[ y_\mu = s^0 + \epsilon_\mu. \]  

(367)

We would like to estimate \( s^0 \) from the measurements \( y_\mu \). To this end we apply Bayes rule to compute a posterior distribution

\[ P(s|y) \propto \prod_{\mu=1}^N P_s(y_\mu - s)P_s(s) = e^{\sum_{\mu=1}^N \log(P_s(y_\mu - s))}P_s(s) = e^{\sum_{\mu=1}^N \log(P_s(\epsilon_\mu))}P_s(s). \]  

(368)

For large \( N \), the posterior distribution will be concentrated in the vicinity of \( s^0 \). Therefore we can expand the distribution around \( s^0 \), taking \( s - s^0 \) to be small:

\[ \sum_{\mu=1}^N \log \left( P_s(\epsilon_\mu - (s - s^0)) \right) \approx \sum_{\mu=1}^N \log' \left( P_s(\epsilon_\mu) \right)(s - s^0) + \frac{1}{2} \sum_{\mu=1}^N \log'' \left( P_s(\epsilon_\mu) \right)(s - s^0)^2 \]  

(369)

These Taylor series coefficients are random variables and as \( N \to \infty \) the central limit theorem implies:

\[ \sum_{\mu=1}^N \log' \left( P_s(\epsilon_\mu) \right) \rightarrow N(0, NJ[\epsilon]), \]  

(370)

\[ \sum_{\mu=1}^N \log'' \left( P_s(\epsilon_\mu) \right) \rightarrow NJ[\epsilon]. \]  

(371)

Thus, in the large \( N \) limit, the precise measurement vector \( y \) in the prior distribution (368) may be replaced with random variables so that

\[ P(s|s^0, z) \propto P_s(s) e^{-\frac{(s - s^0 - z)^2}{2NJ[\epsilon]}}, \]  

(372)

where \( z \) is a unit gaussian random variable and we define \( q_d = \frac{1}{NJ[\epsilon]} \), so that \( q_d \) is the unbiased error from the Classical Cramer-Rao bound, achieved by maximum likelihood in the large \( N \) limit.

\[ \hat{s}^{\text{MMSE}} = \int s \rho_s(s|s^0, z) ds. \]  

(373)

It also follows that the MMSE achievable given the data is simply

\[ q_s = \langle \langle \int (s - \hat{s}^{\text{MMSE}})^2 \rho_s(s|s^0, z) ds \rangle \rangle_{s, z}. \]  

(374)

Using the relationship between MMSE and Fisher information in Appendix B.4, the MMSE of inferring a random variable \( s^0 \) contaminated by gaussian noise of variance \( q_d \), may be written in terms of the Fisher information:

\[ q_s = q_s^{\text{MMSE}}(q_d) = q_d - q_d^2 J \left[ s^0 + \sqrt{q_d} z \right]. \]  

(375)

Then, the convolution Fisher inequality (171) implies

\[ \frac{1}{J \left[ s^0 + \sqrt{q_d} z \right]} \geq \frac{1}{J \left[ s^0 \right]} + q_d, \]  

(376)
so that

$$q_s = q_d - q_d^2 J \left[ s^0 + \sqrt{q_d} z \right] \geq q_d \left( 1 - \frac{q_d}{J \left[ s^0 \right]} + q_d \right) = q_d \left( \frac{1}{1 + q_d J \left[ s^0 \right]} \right) = \frac{1}{\frac{1}{q_d} + J \left[ s^0 \right]} = \frac{1}{NJ \left[ \epsilon \right] + J \left[ s^0 \right]}, \quad(377)$$

yielding a Fisher information bound on asymptotic MMSE inference for a single parameter:

$$q_s \geq \frac{1}{NJ \left[ \epsilon \right] + J \left[ s^0 \right]}.$$ \hfill (378)

Special case of gaussian noise and parameters:

In the case of gaussian noise and parameters ($\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2), s^0 \sim \mathcal{N}(0, \sigma_s^2)$), (378) holds with equality for all $N$ (not just in the asymptotic limit):

$$q_s = \frac{1}{NJ \left[ \epsilon \right] + J \left[ s^0 \right]}.$$ \hfill (379)

Note that $J \left[ \epsilon \right] = \frac{1}{\sigma_\epsilon^2}, J \left[ s^0 \right] = \frac{1}{\sigma_s^2}$, and that we need not consider derivatives higher than second order in $\log(P_\epsilon(x))$, so we need only compute two terms: (370) and (371). In both cases the equalities hold exactly, since a sum of gaussian random variables is a gaussian random variable. Since we do not need to invoke the central limit theorem, this result holds for any $N$. Further, the convolutional Fisher inequality (376) becomes an equality for gaussian, this follows from that fact that the Fisher information is the inverse variance for a gaussian:

$$\frac{1}{J \left[ s^0 + \sqrt{q_d} z \right]} = \frac{1}{\sqrt{J \left[ q_d + \sigma_s^2 z \right]}} = q_d + \sigma_s^2 = q_d + \frac{1}{J \left[ s^0 \right]}.$$ \hfill (380)

### B Fisher Information

#### B.1 Fisher information for scalar inference

The scalar Fisher information $J \left[ \epsilon \right]$ quantifies the accuracy of any unbiased estimator to reconstruct a signal corrupted by additive noise $\epsilon$ drawn from probability density function $P_\epsilon$. The problem is to estimate the scalar $s^0$ given a corrupted measurements $y$:

$$y = s^0 + \epsilon.$$ \hfill (381)

The definition of Fisher information in this scalar case is:

$$J \left[ \epsilon \right] = \left\langle \left( \frac{d}{dx} \log P(y|s^0) \right)^2 \right\rangle_\epsilon = \int \frac{P_\epsilon'(z)^2}{P_\epsilon(z)} dz.$$ \hfill (382)

#### B.2 Convolutional Fisher inequality

The scalar Fisher information obeys the following inequality:

$$J \left[ \epsilon_1 + \epsilon_2 \right] \leq \frac{J \left[ \epsilon_1 \right] J \left[ \epsilon_2 \right]}{J \left[ \epsilon_1 \right] + J \left[ \epsilon_2 \right]},$$ \hfill (383)

where $\epsilon_1, \epsilon_2$ are drawn from separate probability densities. See [6] p.674.

#### B.3 Fisher information is minimized by a gaussian distribution

For a noise distribution $P_\epsilon$, the Fisher information is bounded below as

$$J \left[ \epsilon \right] \geq \frac{1}{\sigma_\epsilon^2},$$ \hfill (384)

and for a fixed variance $\sigma_\epsilon^2$ is minimized for a gaussian distribution. This result follows from the inequality:

$$1 = \left( \int P_\epsilon(x) \left( \frac{(x - \langle \epsilon \rangle)P_\epsilon'(x)}{P_\epsilon(x)} \right) dx \right)^2 \leq (x - \langle \epsilon \rangle)^2 P_\epsilon(x) dx \int \frac{P_\epsilon'(x)^2}{P_\epsilon(x)} dx = \sigma_\epsilon^2 J \left[ \epsilon \right],$$ \hfill (385)
where the left equality follows from integration by parts, the inequality follows from Cauchy Schwartz, and the Fisher information appears in the final equality by substituting the definition of Fisher Information (382). This proves the desired bound. Moreover, the cauchy-schwartz that the bound is saturated when
\[ P'(x) \propto (x - \langle \epsilon \rangle)^2. \] (386)
Integrating both sides and choosing appropriate constants s.t. \( P' \) in a probability distribution yields
\[ P(x) \propto e^{-\frac{(x - \langle \epsilon \rangle)^2}{2\sigma^2}}, \] (387)
so that this bound is saturated only for a gaussian \( P \).

### B.4 Relation between MMSE and Fisher information

The MMSE of estimating a single random variable \( s^0 \) from measurements corrupted by gaussian noise of variance \( q_d \) can be written in terms of the Fisher Information as
\[
q_{s_{\text{MMSE}}}^2(q_d) = q_d - q_d^2 J \left[ s^0 + \sqrt{q_d}z \right].
\] (388)
See e.g. [7] equation 28, p. 38.

### B.5 Cramer-Rao bound

For the case of multiple parameters, the Cramer-Rao bound is written in terms of the Fisher information matrix, we will state this bound here and derive the form of the Fisher information matrix for the problem considered in this work.

For any estimator \( \hat{s} \) of the true parameters \( s^0 \) given the data \( z \), where the conditional probability of data given parameters is \( p(z|s^0) \), we define the Fisher information matrix as
\[
I_{ij} = \left\langle \left\langle \frac{\partial}{\partial s^i} \log p(X, y|s^0) \frac{\partial}{\partial s^j} \log p(X, y|s^0) \right\rangle \right\rangle_z.
\] (389)
The Cramer-Rao bound states that for any unbiased inference algorithm with estimates \( \hat{s} \), so that
\[
\left\langle \left\langle (\hat{s}_k - s^0_k)^2 \right\rangle \right\rangle_z \geq I^{-1}_{kk}.
\] (390)

**Applying Cramer-Rao to the problem considered in this paper:**

We first derive the form of the Fisher information matrix \( I \) for the system studied in this paper:
\[
y = Xs^0 + \epsilon,
\] (391)
where \( \epsilon \sim f \) iid, \( y \in \mathbb{R}^N, s^0 \in \mathbb{R}^P \), and \( X \in \mathbb{R}^{N \times P} \). We apply the definition of the Fisher information matrix (389) to derive its form:
\[
I_{jk} = \left\langle \left\langle \frac{\partial}{\partial x_j} \log p(X, y|s^0) \frac{\partial}{\partial x_k} \log p(X, y|s^0) \right\rangle \right\rangle_{X,y}.
\] (392)
In this problem,
\[
\log p(X, y|s^0) = \log p(y|X, s) + \log p(X) = \sum_{\mu=1}^N \log f(y_{\mu} - X_{\mu} \cdot s) + \log p(X),
\] (393)
where \( X_{\mu} \) is the \( \mu \)th row of the matrix \( X \). Differentiating (392) yields
\[
I_{jk} = \left\langle \left\langle \sum_{\mu=1}^N X_{\mu j} X_{\mu k} \left( \frac{f'(\epsilon_{\mu})}{f(\epsilon_{\mu})} \right)^2 \right\rangle \right\rangle_{X,\epsilon}.
\] (394)
For \( j \neq k \) the input matrices will be uncorrelated so \( I_{jk} = 0 \). For \( j = k \) then,
\[
I_{jj} = \alpha \int dx \frac{f'(x)^2}{f(x)}.
\] (395)
Thus, the Fisher information matrix has no dependence on \( s^0 \) and has the form:
\[ I_{ij} = \delta_{ij} \alpha \int \frac{(f'(z))^2}{f(z)} \, dz = \delta_{ij} \alpha J[\epsilon]. \]  

Substituting the Fisher information matrix into the Cramer-Rao bound, yields the variance per parameter:

\[ q_x \geq \frac{1}{\alpha J[\epsilon]}. \]  

## C Moreau envelope and proximal map

### C.1 Relation between proximal map and Moreau envelope

The Moreau envelope is a functional mapping parameterized by a regularization scalar \( \lambda \). It maps a function \( f \) to

\[ \mathcal{M}_\lambda[f](x) = \min_y \left[ \frac{(x-y)^2}{2\lambda} + f(y) \right]. \]

We will denote the special case of \( \mathcal{M}_1[f] \) by \( \mathcal{M}[f] \). Some properties that follows from this definition are that the minimizers of \( f \) and \( \mathcal{M}_\lambda[f] \) are the same, and that the Moreau envelope is also lower bound on the function \( f \). The closely related proximal map is defined as

\[ \mathcal{P}_\lambda[f](x) = \arg \min_y \left[ \frac{(x-y)^2}{2\lambda} + f(y) \right]. \]

The proximal map can be viewed as a gradient descent step along the Moreau envelope:

\[ \mathcal{P}_\lambda[f](x) - x = -\lambda \mathcal{M}_\lambda[f]'(x). \]

To derive (400), we begin with the derivative of the Moreau envelope and show

\[ \mathcal{M}_\lambda[f]'(x) = \frac{d}{dx} \min_y \left[ \frac{(x-y)^2}{2\lambda} + f(y) \right] = \frac{d}{dx} \left[ \frac{(x-\hat{y})^2}{2\lambda} + f(\hat{y}) \right], \]

where \( \hat{y} \) is the minimizer of the RHS argument of (401). Differentiating with respect to \( \hat{y} \) yields 0 at the minimum, so the \( \hat{y} \) may be effectively treated as a constant and we need only differentiate with respect to \( x \), which yields

\[ \mathcal{M}_\lambda[f]'(x) = \frac{x - \hat{y}}{\lambda}. \]

It follows that

\[ \mathcal{M}_\lambda[f]'(x) = \frac{x - \mathcal{P}_\lambda[f](x)}{\lambda}. \]

### C.2 Inverse of the Moreau envelope

**Result 23. Procedure for inverting a Moreau envelope:** For \( \lambda > 0 \) and \( f \) a convex, lower semi-continuous function such that \( \mathcal{M}_\lambda[f] = g \), the Moreau envelope can be inverted so that \( f = -\mathcal{M}_\lambda[-g] \).

To derive this result, we first consider the case of \( \lambda = 1 \), from which the \( \lambda > 0 \) case will follow. Our assumption that \( \mathcal{M}_\lambda[f] = g \) implies

\[ g(x) = \mathcal{M}[f](x) = \min_y \left[ \frac{(x-y)^2}{2} + f(y) \right] = \frac{x^2}{2} + \min_y \left[ -xy + \frac{y^2}{2} + f(y) \right] = \frac{x^2}{2} - \max_y \left[ xy - \frac{y^2}{2} - f(y) \right]. \]

We now define the Fenchel conjugate \( ^* \), which operates on a function \( h \) to yield \( h^*(x) = \max_y [xy - h(y)] \). We then define, for notational simplicity, the function \( p_2(x) = \frac{x^2}{2} \). With this notation, (404) reduces to

\[ g = p_2 - (f + p_2)^* \]

The Fenchel-Moreau theorem [8] states that if \( h \) is a convex and lower semi-continuous function, then \( h = (h^*)^* \). These properties are assumed true for \( f \) and will also hold for \( f + p_2 \) so that (405) may be inverted, yielding:

\[ f = (p_2 - g)^* - p_2. \]
We now write $f$ in terms of a Moreau envelope by expanding the previous expression:

$$f(x) = -\frac{x^2}{2} + \max_y (xy - \frac{y^2}{2} + g(y)) = -\min_y \left[ \frac{(x-y)^2}{2} - g(y) \right] = -\mathcal{M}[-g](x). \tag{407}$$

Thus, $\mathcal{M}[f] = g$ implies $f = -\mathcal{M}[-g]$. To extend this to $\lambda \neq 1$, we use the identity

$$\lambda \mathcal{M}_\lambda[\frac{1}{\lambda} f] = \mathcal{M}[f], \tag{408}$$

which can be verified by substitution into the definition of the Moreau envelope (398). Combining the result $\mathcal{M}[f] = g$ implies $f = -\mathcal{M}[-g]$ with (408) yields that:

$$\mathcal{M}_\lambda[\frac{1}{\lambda} f] = \frac{1}{\lambda} g, \tag{409}$$

also implies

$$\frac{1}{\lambda} f = -\mathcal{M}_\lambda[-\frac{1}{\lambda} g], \tag{410}$$

which completes the derivation since $\frac{1}{\lambda}$ may be absorbed into the definition of $f$ and $g$.

### C.3 Conditions for convexity of optimal M-estimators

A sufficient condition to guarantee the convexity of an optimal regularized M-estimator is log-concavity of the signal and noise distributions $P_s$ and $P_r$. Here we explain why this is the case. The main reason is that the optimal loss and regularizer should be convex to guarantee the validity of the RS approximation in the zero temperature replica theory for M-estimation. First, we note that the functional relationship between the optimal loss (or regularizer) and the noise (or signal) distribution should be convex to guarantee the validity of the RS approximation in the zero temperature replica theory for M-estimation.

In terms of a Moreau envelope by expanding the previous expression:

$$h^{\text{opt}} = -\mathcal{M}_q[\log f_q](x) = (q \log f_q + p_2)^* - p_2, \tag{411}$$

where the function $f$ denotes the noise (or signal) distribution, the function $h^{\text{opt}}$ denotes the optimal loss (or regularizer), and $q$ denotes the level of Gaussian smoothing. Then the question we must ask is, under what conditions on $f$ is $h^{\text{opt}}$ convex?

The basic idea is that if $f$ is a log-concave distribution, then $f_q$ is the convolution of a pair of log concave distributions, which will also be log concave; i.e. $\log f_q$ is a concave function. Now the sum $q \log f_q + p_2$ is the sum of a concave function with the convex function $p_2(x) = \frac{1}{2} x^2$. However, it can be shown that this sum is convex for any $q$. Moreover, this sum is in a sense less convex than $p_2$ itself (a function $a$ is more convex than $b$ if $a - b$ is convex). Furthermore the Fenchel conjugate of a function less convex than $p_2$ is more convex than $p_2$. Thus $(q \log f_q + p_2)^*$ is more convex than $p_2$. Therefore, subtracting the convex function $p_2$ from it, does not destroy its convexity. Therefore $h^{\text{opt}}$ is convex. See [2] for a more detailed derivation of this result. Finally, we stress that log-concavity of the signal and noise is a sufficient condition for the validity of our optimal M-estimators, but it may not be necessary.

### References


