Excerpts from students' first drafts

These excerpts are drawn from the first drafts of your reports. For each excerpt, suggest ways to improve the use of mathematics.

?　　　introduce.　　　a measurement description

1. In this work, we assume a measurement of the signal containing N pixels, $x \in \mathbb{R}^N$ is corrupted by additive noise, η: $y = x + \eta$, where $y \in \mathbb{R}^N$ is the noisy signal and we often assume that η follows $\mathcal{N}(0, \sigma^2 I_N)$. The denoising problem involves finding a f such that the noisy observation y can be mapped to a good estimate of x, i.e, $f(y) \approx x$. Often, one says that $f(y)$ is as good denoised image if the squared error, $||x - f(y)||^2$ is small. Thus, $f = \arg\min_g E||x - g(y)||^2$, where the expectation is taken over some distribution over images, x, as well as over the distribution of noise realizations. If the denoiser $f(.)$ is expected to work for a range of noise levels σ, then the problem is called blind denoising. In this case, the expectation has to taken over σ as well.

   - f - multiple notations for it.
   - argmin over what? What is g? g underneath
   - $\mathbb{E}$ over what
   - σ is not defined, I_N.
   - What is g?

2. The main part of the generalized Debye framework that must be modified to be able to be able to work with non-smooth objects are the steps that requires solving surface Poisson problems, i.e. problems of the form

   define.　　　　refer to later?

$$\Delta_\Gamma u = f, \tag{1}$$

   where Δ_Γ is the surface Laplacian for the surface Γ. As the surface Laplacian is only defined in regions where the surface is smooth, some modification must be made to incorporate the edges and corners in the surface. Previous work by Chernokozhin and Boag [1] suggests that this modification can be done by replacing the surface Poisson equation with surface Poisson equations on each face of the surface Γ_i and matching conditions at the edge between faces i and j, e_{ij}. These requirements may be written as　　long

   more space,

$$\Delta_{\Gamma_i} u = f \ \forall i, \text{ and } \ u|_{e^+_{ij}} = u|_{e^-_{ij}} \ \text{ and } \ \left.\frac{du}{d\nu}\right|_{e^+_{ij}} = \left.\frac{du}{d\nu}\right|_{e^-_{ij}} \ \forall i, j, \tag{2}$$

   where ν is the vector tangent to the surface but normal to the edge. While (2) takes roughly the form I expect based on intuition gained from similar problems and the numerical results given by Chernokozhin and Boag are promising, the assumptions they used to derive (2) are not obviously satisfied. As such, I am working to derive (2) from first principles if possible and from numerical evidence if it is not.

   what derive?

3. Homogenization theory is a useful tool to draw the connection between macroscopic and microscopic behavior of a periodically oscillatory problem. Mathematically, we would like to solve the problem $div(a(x/\epsilon)\nabla u^\epsilon(x)) = f(x)$ in $H_0^1(\Omega)$ with $a(y)$ Q-periodic and $\epsilon \to 0$. And u^ϵ weakly converges to $u^*(x)$, with $u^*(x)$ solves $div(a^*_{(x)}\nabla u^*(x)) = f(x)$ with a^* as the effective tensor.

DISPLAY

- solve — for some a in Sobolev space H^1 etc
- $\varepsilon > 0$, then $\varepsilon \to 0$
- define a earlier
- what is ε physically?
- What is a(x) in the context of this problem?

4. We propose a novel top-down multi-layer convolutional architecture trained to produce a hierarchy of sparse overcomplete representations of a given image input. It consists of a convolutional decoder $\mathcal{D}$ which takes as input a set of codes at different scales $Z = \{Z_i\}_{i=0}^{k-1}$ to generate a reconstruction $\hat{y}$ of a given input y. [...] Each code $Z_i \in \mathbb{R}^{c_i \times w_i \times h_i}$ is a collection of c_i feature maps of dimension $w_i \times h_i$. [...]

   Finding the optimal sparse codes for a fixed decoder $\mathcal{D}$ is done by taking multiple steps of the fast iterative shrinkage-thresholding algorithm (FISTA) proposed in [1]. FISTA is a version of the iterative shrinkage-thresholding algorithm (ISTA) which speeds up convergence. In the context of our model, a regular ISTA update of code Z_i is computed as follows:

$$Z_{i,t+1} = \tau_{\alpha_i \eta_i}(Z_{i,t} - \eta_i \nabla_{Z_{i,t}} \text{MSE}(y, \hat{y})), \tag{3}$$

- can define more simply.

   where t is the time step, α_i is a sparsity coefficient, η_i is the step size, and $\tau_\beta \colon \mathbb{R}^n \to \mathbb{R}^n$ is the shrinkage function:

$$\tau_\beta(x)_i = \text{sign}(x_i)(|x_i| - \beta)_+. \tag{4}$$

used in diff't ways

   The shrinkage function forces components of its input to become 0 if their absolute value is less than β. FISTA uses the same update as ISTA (3) but instead of $Z_{i,t}$, it is applied to a linear combination of the past two iterates $Z_{i,t}$ and $Z_{i,t-1}$.

- 1st ¶ is confusing, needs diagram.
- two subscripts in Z, τ. Can simplify?
- MSE not defined
- τ_β defined earlier
- $\nabla_{Z\ldots}$ MSE – what does ∇_Z mean?
- Subscript of τ is multiplication – confusing.
- English descr. 1st of (3), (4).

5. Gaussian Processes have different definitions depending on the level of abstraction of the field. Here, we consider Gaussian Processes from the perspective of statistics and machine learning. Roughly, this interpretation consists on imposing a Gaussian prior (or belief) over the space of functions. More precisely, let $f(x)$ be a function from $\mathbb{R}^n$ to $\mathbb{R}$ and let $\mathcal{D}_n = \{(x_i, f(x_i)), i \in 1..n\}$ be a dataset consisting of pairs of evaluation. We say that *** is a Gaussian process over f given $\mathcal{D}_n$ if the probability for all $x \in \mathbb{R}^n$ is a Gaussian that is $p(f(x)|\mathcal{D}_n) \sim \mathcal{N}(\mu_{f|\mathcal{D}_n}(x), k_{f|\mathcal{D}_n}(x, x'))$ with $\mu_{f|\mathcal{D}_n}(x)$ the mean function on a particular point x and $k(x, x')$ the correlation function between x and x'. Intuitively, the former definition means that at each point x we have a Gaussian distribution over the possible images $f(x)$ with parameters that depend on the dataset and the point x.

Good!

Go earlier!

- bold for vectors & regular for scalars (?)
- restructure: intuition 1st, then display [ ]
- How do f, D_n enter in def on μ, k.
- $D_n \to D$ (?)

6. . . . [1 paragraph non-mathematical introduction] . . .
First, we aim to find a metric describing how well a classifier performs. For simplicity, we will discuss classification into two classes, denoted −1 and +1. Our classifiers are then measurable functions $f : \mathcal{X} \to \{-1, +1\}$. The *classification error* of a classifier f is

not defined

what are x, y, D?

$$R(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\mathbf{1}_{f(x)\neq y}] \tag{5}$$

needed? D doesn't change

where our data is distributed according to a distribution $\mathcal{D}$.

One common choice for f is to choose $g : \mathcal{X} \to \mathbb{R}$ and then set f to match the sign of g:

$$f(x) = \operatorname{sign} g(x) = \begin{cases} 1 & \text{if } g(x) > 0 \\ -1 & \text{if } g(x) \le 0 \end{cases}$$

Why g first, then f?

necessary?

(The sign of 0 was chosen arbitrarily.) The decision boundary is then $\{x : g(x) = 0\}$. The value $|g(x)|$ can then be interpreted the confidence in our prediction. It may seem reasonable to consider $\mathbf{1}_{g(x)y<0}$ rather than $\operatorname{sign} g(x)$. Note that the function $\mathbf{1}_{\operatorname{sign}(g(x))\neq y}$ does not depend on $|g(x)|$ at all. For correctly classified x, it would then seem reasonable to penalize large values of $|g(x)|$ less than small values of $g(x)$. For this end, we replace the indicator function in (5) with a decreasing function ϕ and consider

necessary?

$$\mathcal{L}_\phi = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\phi(yg(x))] \tag{6}$$

Note that if if $\operatorname{sign} g(x)$ correctly classifies x, then $yg(x) \ge 0$. The quantity $\phi(yg(x))$ is often used in algorithms instead of $\mathbf{1}_{\operatorname{sign} g(x)\neq y}$ for ease of computation. It's common to choose a ϕ with $\phi(z) \ge \mathbf{1}_{z<0}$, and then ϕ is called a *surrogate loss for* $\mathbf{1}_{z<0}$.

- dense. Nice to start w/ example.
- Give intuition for f.
- Why is notation for R(f), L_ϕ so different
- ~~too~~ technical, start w/ concrete example.